

京都大学
情報学研究科
博士論文
Kyoto University
Graduate school of Informatics
Doctoral thesis

視覚における知覚体制化のための時間相関の計算機構
Computing Temporal Correlation for Perceptual Organization in Vision

陳 彦儒
Yen-Ju Chen
指導教員：西田 眞也
Supervisor: Shin'ya Nishida

令和 7 年 9 月
September 2025

Declaration

I, Yen-Ju Chen, declare that this thesis, entitled “Computational characteristics of temporal correlation in perceptual organization: mechanism and system dynamics” is original and my own work. I confirm that:

- This work was done solely while I was a candidate for the research degree at Graduate School of Informatics, Kyoto University.
- No part of this work has previously been submitted for a degree at this or any other university.
- References to the works of others have been clearly attributed. Questions from the work of others have been clearly indicated and attributed to them.
- In cases where others have contributed to parts of this work, such contributions have been clearly acknowledged and distinguished from my own work.

Yen-Ju Chen

07 2025

Abstract

This thesis investigates the human visual system's capacity to perform segmentation based on dynamic temporal information, with a particular emphasis on whether such segmentation reflects an attribute-invariant computational process. While previous models have made significant progress in explaining spatial grouping based on low-level visual features, they often fail to generalize segmentation tasks involving purely temporal or cross-attribute cues. To address this gap, the present work adopts a systematic psychophysical and computational modeling approach to elucidate the mechanisms underlying temporal segmentation, characterize its invariance properties, and clarify the associated temporal dynamics.

In Chapter 1, I first review the theoretical and computational background of perceptual segmentation, including key frameworks such as texture statistics, filter-rectify-filter (FRF) model. While these models provide valuable insights into spatial grouping, their limitations become evident when confronted with time-varying stimuli and cross-attribute interactions. I then motivate the need to treat temporal segmentation as a distinct computational challenge and outline four central research questions: (1) Does temporal segmentation operate independently of specific visual attributes? (2) How do signals from different attributes interact during temporal segmentation? (3) What kind of computation underlies this attribute-invariant process? (4) What are the temporal dynamics and integration profiles of the underlying mechanism?

Chapter 2 provides the theoretical and methodological scaffolding for the experiments that follow. I begin with an analysis of perceptual organization within the visual hierarchy and its neural basis, highlighting how grouping supports object-based representation and ambiguity reduction. I then formalize temporal processing using the linear time-invariant (LTI) system framework and introduce its application to visual impulse response and system transfer functions. I also present the algebraic foundations of graph theory as applied to segmentation and introduce the continuous tracking paradigm with cross-correlogram (CCG) analysis as a method for probing system latency and dynamic response.

In Chapter 3, I directly test the existence of attribute-invariant temporal segmentation. I extend Guttman et al (2005) method by using temporal modulations across luminance, color, and motion direction, as well as waveform type. Results show that segmentation tuning curves exhibit similar frequency profiles across all attributes, with peak sensitivity at 1–5 Hz, provided that temporal transiency is preserved. Both within- and cross-attribute conditions show comparable performance, providing strong evidence that the temporal segmentation process generalizes across visual features and is not strictly tied to specific stimulus dimensions.

In Chapter 4, I examine the interaction between visual attributes during segmentation. Using noise-masked stimuli that interfere with temporal correlation signals, I compare the effects of noise from the same vs. different feature channels. Results show that noise from different attributes impairs performance comparably to noise from the same attribute, suggesting that signal integration precedes attribute-independent processing. A computational model based on linear summation of gain-weighted signals from different channels captures the majority of observed variance, providing further support for a unified, amodal segmentation mechanism.

Chapter 5 explores the computational principles underlying this invariance. A novel stimulus generation method is introduced based on a Vision Transformer (ViT) architecture, which creates stimuli with equated low-level statistics and controlled temporal correlation structures while minimizing spatial cues. Human behavioral data reveal that segmentation performance is directly modulated by the magnitude of temporal correlation. Motivated by this, I construct a graph-based segmentation model using Laplacian decomposition, which treats temporal correlation as an edge weight and applies spectral clustering for segmentation prediction. Model outputs align closely with human data, supporting the notion that temporal segmentation may be instantiated through image-computable, graph-based representations.

Chapter 6 shifts focus to the temporal dynamics of the segmentation process by introducing a continuous tracking paradigm designed to dissociate perceptual processing from motor response. Two task variants were developed: position tracking (PT) and attribute tracking (AT). Their temporal response profiles were compared using cross-correlogram (CCG) analysis. Results revealed similar core processing latencies across tasks, but AT exhibited an additional delay attributable to its higher mapping complexity. A follow-up experiment demonstrated that PT performance could be equivalently slowed by artificially increasing both the mapping complexity and signal magnitude, thereby confirming that PT may serve as a reliable proxy for estimating perceptual latency under controlled conditions. These findings establish continuous tracking as a sensitive and flexible method for probing temporal dynamics in the visual system, while also highlighting systematic differences between spatially grounded and feature-based processing streams.

Chapter 7 integrates the reaction-time (RT) task with the continuous tracking paradigm to examine the temporal integration window underlying correlation-based segmentation. Crucially, a computational model is developed to decompose the observed position tracking (PT) behavior into components that are specifically related to correlation-based processing and those that are not. This dissociation allows us to isolate the portion of the system's temporal dynamics that directly reflects correlation integration. By jointly analyzing RT and tracking data under varying levels of temporal correlation, the model estimates that the effective integration window spans approximately 600 to 1000 milliseconds. Furthermore, this window shows remarkable consistency across different visual attributes, reinforcing the hypothesis that a shared, feature-invariant mechanism governs temporal segmentation.

Chapter 8 provides a general discussion that integrates the main findings and situates them within broader theories of perceptual organization. The results support the existence of a feature-invariant temporal segmentation mechanism based on pairwise temporal correlation, represented and computed through a graph-like structure. The system integrates information over extended time windows and shows consistent behavior across visual attributes. A gain-based summation model accounts for cross-attribute

signal merging prior to segmentation, offering a parsimonious explanation for the observed invariance. Open questions remain regarding the generalizability to spatial segmentation, sensitivity to complex shapes, and the neural implementation of graph-like processing.

In summary, this thesis presents converging psychophysical and computational evidence that temporal segmentation in human vision is governed by a generalized, attribute-invariant mechanism operating over long integration windows. The process is consistent across luminance, color, and motion domains, sensitive to waveform variation, and explainable through a graph-based spectral model. Together, these results help formalize temporal segmentation as a mid-level visual computation and offer a new foundation for future research on dynamic perceptual organization. The methodology developed here—including continuous tracking, ViT-based stimulus design, and graph-based modeling—may also serve as a bridge toward human-aligned artificial vision systems capable of generalizing across sensory features and temporal contexts.

Acknowledgement

A summer noon in Kyoto: thick heat hangs like steam and seeps into my bones, draining both strength and thought. I sit in front a blank screen, writing a line only to erase it, then starting again, still unable to catch the right opening for the close of three years' journey. People say the acknowledgement is easy—drop a few names, turn the page—yet none of the people I name here is just a passing stranger in my life. It is this pull between honest gratitude and tidy academic form that keeps my fingers hovering, looking for the “perfect” full stop. That same tension has colored every stage of my studies since college, and I suspect it will follow me as long as I keep chasing a theory that feels true.

First in this story is my advisor, 西田先生. Strict and exact, he broke my early ideas apart and pushed me to rebuild them with cleaner lines. Our meetings were rarely gentle, but they trained me to hold on to a thread of logic in the rush of talk and to speak that logic without fear. Over time I even learned to enjoy the heat of those debates. In our field, that taste for sharp argument is a skill worth keeping. My feelings toward 西田先生 mix respect, irritation, and some complicated feeling under both, probably can be named as gratitude.

Next are the friends and colleagues who rode every wave of my moods. 孫子崇, partner in one scheme after another, listened to my wild or half-built ideas and answered each with care until many turned into real results. 李桑, 何桑, **Jason** fellow PhD travelers, often pulled this indoor hermit outside, adding memories that had nothing to do with data or deadlines. Also, Great gratitude 詠浩, 三好桑 and 大久保桑 for always have a discussion and inspired my chaotic thought.

The circle widens: master's students 姜桑, 朱桑, 今村, 田中, 平野, 大矢, whose questions filled the lab with active and warm atmosphere; the Peking University friends I kept meeting at VSS—彦儒, 彩霞, 王馨; the lifelong friends in Taiwan always stand

by side in all my climax and valley—宗鴻, 高欣, 曾元, 建華, and of course 聰宇; **Professor Johnston** and **Akira** in Nottingham, who offered guidance and honest debate; and my oral-exam committee, 神谷先生 and 蘆田先生, whose constructive advisement during the defense still echo in these pages.

And always, my family—father, mother, two elder sisters, and our dog 嚕嚕米—whose quiet faith steadied every uncertain step.

I began determined not to write a dull list of thanks, yet here the names flow and with each one I feel lighter. So be it. Human nature is full of small clashes, and bright ideas often rise from them. Every good person, good moment, and stroke of luck is a treasure I will count again and again. The road so far has been kind, even when rough. I carry these gifts forward as I set out on the next stretch of this wide sea of learning.

To all the people and moments that shaped this journey—thank you.

Contents

DECLARATION	I
ABSTRACT	III
ACKNOWLEDGEMENT	VII
CONTENTS	IX
LIST OF FIGURES.....	XIV
LIST OF TABLES	XVI
LIST OF PUBLICATION.....	XVII
CHAPTER 1. GENERAL INTRODUCTION	19
1.1. COMPUTATIONAL APPROACH OF PERCEPTUAL GROUPING/SEGMENTATION.....	20
1.2. PRIMARY LIMITATION OF CURRENT COMPUTATIONAL APPROACH	23
1.3. DYNAMIC CUE FOR PERCEPTUAL GROUPING/SEGMENTATION	24
1.4. RESEARCH MOTIVATION AND SCAFFOLDING	32
CHAPTER 2. THEORETICAL BACKGROUND.....	36
2.1. THE FUNCTIONAL ROLE OF PERCEPTUAL ORGANIZATION IN VISION HIERARCHY.....	37
2.2. NEURAL BASIS FOR PERCEPTUAL GROUPING/SEGMENTATION	39
2.3. TEMPORAL PROCESSING IN THE VISUAL SYSTEM	42
2.3.1. <i>Linear time invariant system approach</i>	42
2.3.1.1. Impulse response and convolution	43
2.3.1.2. Eigenfunction property of LTI system.....	44
2.3.1.3. Duality and Relationship Between Impulse Response and Transfer Function	45
2.3.2. <i>Impulse response measured by classical method</i>	46
2.3.2.1. Temporal-domain-based approach	46
2.3.2.2. Frequency-domain-based approach.....	47
2.4. GRAPH THEORY IN PERCEPTUAL ORGANIZATION	49
2.4.1. <i>Algebraic graph basis of organization task</i>	49
2.4.2. <i>Relevance to perceptual grouping models</i>	50
2.4.2.1. Normalized cut and graph spectrum.....	51
2.4.2.2. Attention-Based Deep Neural Networks.....	54
2.5. CONTINUOUS TRACKING PARADIGM AND CROSS-CORRELOGRAM.....	56
CHAPTER 3. FEATURE INVARIANT PROCESS OF TEMPORAL SEGMENTATION.....	60
3.1. INTRODUCTION	60
3.2. EXPERIMENT 1-1. VISUAL SEGMENTATION ON LUMINANCE, COLOR, AND MOTION DIRECTION	63
3.2.1. <i>Method</i>	63

3.2.1.1.	Apparatus.....	63
3.2.1.2.	Participants	63
3.2.1.3.	Stimuli	64
3.2.1.4.	Design and Procedure	67
3.2.2.	<i>Result</i>	68
3.2.3.	<i>Interim discussion</i>	71
3.3.	EXPERIMENT 1-2. VISUAL SEGMENTATION ON DIFFERENT WAVEFORMS.....	72
3.3.1.	<i>Method</i>	73
3.3.2.	<i>Result</i>	75
3.4.	GENERAL DISCUSSION.....	80
3.4.1.	<i>Effects of waveform type or the order of changes</i>	80
3.4.2.	<i>Attribute-invariant process</i>	82
3.4.3.	<i>Related Studies</i>	84
3.4.4.	<i>Potential implementation of the attribute invariant computation on neurophysiology</i> ...	85
3.5.	CONCLUSION	86
CHAPTER 4. SINGLE PROCESS CHANNEL UNDERLYING TEMPORAL CORRELATION BASED VISUAL SEGMENTATION		87
4.1.	INTRODUCTION.....	87
4.1.1.	<i>The feature integration issue in temporal segmentation</i>	87
4.1.2.	<i>Temporal correlation as an attribute-invariant processing</i>	88
4.1.3.	<i>Single or multiple channels underlying the process?</i>	88
4.2.	EXPERIMENT 2-1. THE CORRELATION NOISE INTERFERENCE FROM SAME AND DIFFERENT VISUAL ATTRIBUTES	91
4.2.1.	<i>Methodology</i>	91
4.2.1.1.	Apparatus.....	91
4.2.1.2.	Participants	91
4.2.1.3.	Stimuli	92
4.2.1.4.	Redefining the feature tensor	92
4.2.1.5.	Actual stimuli generation.....	93
4.2.1.6.	Design and procedure	94
4.2.1.7.	Analysis	95
4.2.2.	<i>Result</i>	96
4.2.2.1.	Noise disturbance within same attribute channel	96
4.2.2.2.	Noise disturbance across attribute channels	97
4.2.3.	<i>Interim discussion</i>	99
4.3.	EXPERIMENT 2-2. THE INFLUENCE OF LOCAL VARIATION MAGNITUDE	100
4.3.1.	<i>Result</i>	101
4.4.	MODEL CONDUCTION – A SINGLE CHANNEL HYPOTHESIS.....	103
4.4.1.	<i>Model description</i>	104
4.4.2.	<i>Result</i>	105
4.5.	GENERAL DISCUSSION.....	108

4.5.1.	<i>A unified gain-weighted mechanism accounts for cross-attribute temporal segmentation</i>	108
4.5.2.	<i>The integration mechanism needs further research</i>	110
4.5.3.	<i>The neuron implementation of amodal integration</i>	111
CHAPTER 5. FEATURE-INVARIANT SEGMENTATION EMERGES FROM GLOBAL TEMPORAL CORRELATION		112
5.1.	INTRODUCTION	112
5.2.	EXPERIMENT 3-1. STIMULI GENERATION INCORPORATING VIT-BASED GENERATOR	116
5.2.1.	<i>Generation protocol</i>	116
5.2.1.1.	Feature tensor generated by visual transformer	116
5.2.1.2.	Problem definition	116
5.2.1.3.	Correlation constraints	117
5.2.1.4.	Transformer-based generative model	117
5.2.1.5.	Optimization	118
5.2.1.6.	Training procedure	120
5.2.2.	<i>Result</i>	121
5.2.2.1.	Invisibility of spatial cues	121
5.2.2.2.	Validation of cosine similarity distribution	122
5.2.2.3.	Quasi-white noise characteristics	124
5.2.2.4.	Exclusion of higher-order visual feature visibility	125
5.3.	EXPERIMENT 3-2. HUMAN SEGMENTATION PERFORMANCE FOR THE GENERATED STIMULI	125
5.3.1.	<i>Methodology</i>	125
5.3.1.1.	Apparatus	125
5.3.1.1.	Participants	126
5.3.1.2.	Stimuli	126
5.3.1.3.	Design and procedure	127
5.3.2.	<i>Result</i>	128
5.4.	EXPERIMENT 3-3. ANALYSIS OF HUMAN PERFORMANCE USING A GRAPH CUT MODEL	132
5.4.1.	<i>Methodology</i>	134
5.4.1.1.	Model architecture	134
5.4.1.2.	Simulation protocol	137
5.4.2.	<i>Result</i>	138
5.4.2.1.	Naïve model simulation results	138
5.4.2.2.	Full model simulation results	139
5.5.	GENERAL DISCUSSION	142
5.5.1.	<i>Transformer-based generation of complex stimuli</i>	142
5.5.2.	<i>Graph-based representation and its role in mid-level vision</i>	143
5.5.3.	<i>Comparison with CV image segmentation models</i>	145
5.5.4.	<i>Relationship with current neurophysiological evidence</i>	145
CHAPTER 6. A NEW WAY OF MEASURING SYSTEM TEMPORAL DYNAMICS AND ITS CHARACTERISTICS		147

6.1.	INTRODUCTION.....	147
6.2.	EXPERIMENT 4-1. CCG DIFFERENCE BETWEEN PT AND AT	149
6.2.1.	<i>Method</i>	149
6.2.1.1.	Apparatus.....	149
6.2.1.2.	Participants	149
6.2.1.3.	Stimuli	150
6.2.1.4.	Design and Procedure	153
6.2.1.5.	Analysis	154
6.2.2.	<i>Result</i>	155
6.2.3.	<i>Interim discussion</i>	158
6.3.	EXPERIMENT 4-2. S-R MAPPING AND THE EFFECT OF STIMULUS INTENSITY ON POSITION TRACKING.....	160
6.3.1.	<i>Result</i>	161
6.4.	EXPERIMENT 4-3. S-R MAPPING AND LOCAL CONTRAST INFLUENCE ON REACTION TIME	166
6.4.1.	<i>Methodology</i>	168
6.4.2.	<i>Result</i>	169
6.5.	GENERAL DISCUSSION.....	170
6.5.1.	<i>Common attribute encoder for PT and AT</i>	171
6.5.2.	<i>Stimulus-Response mapping</i>	171
6.5.3.	<i>Continuous tracking performance</i>	173
6.5.4.	<i>Potential Model for tracking task</i>	174
6.6.	CONCLUSION	175
CHAPTER 7. TEMPORAL DYNAMICS OF TEMPORAL CORRELATION-BASED VISUAL SEGMENTATION		176
7.1.	INTRODUCTION.....	176
7.2.	EXPERIMENT 5-1. THE REACTION TIME PARADIGM	179
7.2.1.	<i>Methodology</i>	179
7.2.1.1.	Apparatus.....	179
7.2.1.2.	Participants	179
7.2.1.3.	Stimuli	180
7.2.1.4.	Design and procedure	182
7.2.2.	<i>Result</i>	183
7.3.	EXPERIMENT 5-2. THE CONTINUOUS TRACKING PARADIGM.....	184
7.3.1.	<i>Methodology</i>	184
7.3.1.1.	Stimuli	184
7.3.1.2.	Design and procedure	186
7.3.1.3.	Analysis	188
7.3.2.	<i>Result</i>	189
7.4.	COMPUTATIONAL MODEL: INTEGRATING RT AND TRACKING DATA	191
7.4.1.	<i>Model architecture</i>	191
7.4.2.	<i>Result</i>	195

7.5.	GENERAL DISCUSSION	199
7.5.1.	<i>Feature invariant properties of the temporal correlation-modulated process</i>	199
7.5.2.	<i>Limitation</i>	201
7.5.3.	<i>Conclusion</i>	203
CHAPTER 8.	GENERAL DISCUSSION	204
8.1.	FEATURE-INVARIANT PROCESS OF TEMPORAL SEGMENTATION	205
8.2.	INTEGRATED INFORMATION PROCESSING STREAM	208
8.3.	LOW TEMPORAL RESOLUTION AND LONG RECEPTIVE FIELD ON TIME	209
8.4.	SOME INSIGHT INTO THE FUNCTIONAL ROLE OF TEMPORAL SEGMENTATION	209
8.5.	LIMITATION	211
8.5.1.	<i>A more realistic computational model in the natural world</i>	211
8.5.1.1.	Generalizability to pure spatial segmentation and their interaction	212
8.5.1.2.	The relationship with high-order spatial features	213
8.5.1.3.	Multi-attribute integration rules	214
8.5.1.4.	The rule for arbitrary shape extraction/organization	214
8.5.2.	<i>Neurophysiological insight</i>	215
8.6.	CONCLUSION	216
CHAPTER 9.	APPENDIX AND SUPPLEMENTARY MATERIAL	217
9.1.	SUPPLEMENTARY MATERIAL OF CHAPTER 3	217
9.2.	SUPPLEMENTARY MATERIAL OF CHAPTER 5	225
9.3.	SUPPLEMENTARY MATERIAL OF CHAPTER 6	227
REFERENCE		233

List of Figures

FIGURE 1-1. DEMONSTRATION OF 1 ST AND 2 ND ORDER LUMINANCE STATISTICS BASED SEGMENTATION	22
FIGURE 1-2. DEMONSTRATION OF WHETHER VARIOUS SPATIOTEMPORAL TEXTURES CAN BE PARSED USING A SIMPLE FILTER-RECTIFY (FRF) MODEL	28
FIGURE 1-3. CONCEPTUAL FRAMEWORK OF TWO LEVEL OF TEMPORAL SEGMENTATION.	31
FIGURE 2-1. CORTICAL REGIONS INVOLVED IN PERCEPTUAL GROUPING AND SEGMENTATION.	40
FIGURE 3-1. A POSSIBLE COMPUTATIONAL MECHANISM THAT CAN EXPLAIN ATTRIBUTE-INVARIANT SEGMENTATION BASED ON TEMPORAL ASYNCHRONY.	62
FIGURE 3-2. STIMULI FOR EXPERIMENT 1-1.	67
FIGURE 3-3. PROPORTION CORRECT AS A FUNCTION OF THE TEMPORAL FREQUENCY FOR EACH ATTRIBUTE PAIR.	70
FIGURE 3-4. EXPERIMENT 1-2 USED SQUARE AND TRIANGULAR WAVES.	75
FIGURE 3-5. PROPORTION CORRECT AS A FUNCTION OF THE TEMPORAL FREQUENCY FOR EACH ATTRIBUTE AND WAVEFORM PAIR.	78
FIGURE 3-6. THE PROPORTION OF INVERSE RESPONSES AS A FUNCTION OF THE TEMPORAL FREQUENCY OF EACH ATTRIBUTE AND WAVEFORM PAIR.	79
FIGURE 3-7. PEARSON'S CORRELATION COEFFICIENT AS THE FUNCTION OF THE PHASE SHIFT OF A TRIANGULAR WAVE RELATIVE TO A SQUARE WAVE.	82
FIGURE 3-8. TWO PUTATIVE PROCESS FOR THE TEMPORAL CORRELATION-BASED VISUAL SEGMENTATION.	90
FIGURE 3-9. SEGMENTATION PERFORMANCE AS A FUNCTION OF TEMPORAL CORRELATION STRENGTH IN THE WITHIN-ATTRIBUTE CONDITION.	97
FIGURE 3-10. SEGMENTATION PERFORMANCE IN THE CROSS-ATTRIBUTE CONDITION ACROSS ALL COMBINATIONS OF SIGNAL AND NOISE ATTRIBUTES.	98
FIGURE 3-11. DISTANCE INDEX DH FOR ALL CROSS-ATTRIBUTE PAIRINGS	99
FIGURE 3-12. SEGMENTATION PERFORMANCE IN THE CROSS-ATTRIBUTE HALF-NOISE CONDITION OF EXPERIMENT 2-2.	102
FIGURE 3-13. DISTANCE INDEX DH AS A FUNCTION OF LUMINANCE MAGNITUDE IN EXPERIMENT 2-2. .	103
FIGURE 3-14. MODEL PREDICTIONS AND OBSERVED DATA ACROSS ALL EXPERIMENTAL CONDITIONS.	107
FIGURE 5-1. LINKAGE BETWEEN SIMILARITY MATRIX AND SEGMENTATION.	115
FIGURE 5-2. ILLUSTRATION OF THE EXPERIMENT STIMULI.	121
FIGURE 5-3. DISTRIBUTION OF TEMPORAL CORRELATIONS IN THE GENERATED STIMULI.	123
FIGURE 5-4. TEMPORAL SPECTRUM OF THE GENERATED STIMULI.	124
FIGURE 5-5. HUMAN SEGMENTAION PERFORMANCE RESTULT.	131
FIGURE 5-6. MODEL SIMULATION PROTOCOL AND ARCHITECTURE	134
FIGURE 5-7. SIMULATION RESULT FOR THE NAÏVE MODEL.	139
FIGURE 5-8. MODEL OUTPUT OF DIFFERENT PAIRS OF PARAMETERS.	141
FIGURE 5-9. SIMULATION OF DENSE PARAMETER PAIRS AND OPTIMAL MODEL OUTPUT.	142

FIGURE 6-1. STIMULI FOR EXPERIMENT 4-1.	153
FIGURE 6-2 ONE PARTICIPANTS' CCGs FOR PT AND AT ACROSS DIFFERENT ATTRIBUTE TYPES.	156
FIGURE 6-3. CORRELATION BETWEEN PT AND AT.....	157
FIGURE 6-4. PUTATIVE INFORMATION-PROCESSING PIPELINE FOR THE PT AND AT TASKS.	159
FIGURE 6-5 THE CCG OF ONE PARTICIPANT ACROSS THREE PT MODULATIONS AND ATTRIBUTE CONDITIONS.	162
FIGURE 6-6. SCATTER PLOT AND THE STATISTICS ACROSS CONDITIONS.	163
FIGURE 6-7. LATENCY CHANGING RATE FROM THE FIRST 15 TRIALS TO THE LATTER 15 TRIALS ACROSS VARIOUS TRACKING CONDITIONS AND ATTRIBUTE TYPES.	164
FIGURE 6-8. MODEL FIT FOR PEAK LATENCY DATA ACROSS ALL EXPERIMENTAL CONDITIONS.	166
FIGURE 6-9. THE TWO REACTION TIME PARADIGMS WORKFLOW.	168
FIGURE 6-10. EXPERIMENT 4-3 RESULT.....	170
FIGURE 7-1. STIMULUS STRUCTURE AND TRIAL TIMELINE IN THE REACTION TIME TASK.	182
FIGURE 7-2. REACTION TIME AND SELECTION RATE ACROSS LEVELS OF TEMPORAL CORRELATION.....	184
FIGURE 7-3. DEMO SPATIAL AND TEMPORAL WAVEFORM OF THREE CONDITIONS.	187
FIGURE 7-4. CCG RESULT ACROSS PARTICIPANTS ON DIFFERENT ATTRIBUTE AND TRACKING CONDITIONS.	190
FIGURE 7-5. CONCEPTUAL PIPELINE OF DOING TRACKING TASK WITH FIXATED OR DYNAMIC BACKGROUND.	192
FIGURE 7-6. MODEL FITTING RESULT FOR TRACKING TASK FOR EACH PARTICIPATNS AND ATTRIBUTE TYPES	196
FIGURE 7-7. MODEL FITTING RESULT FOR RT TASK FOR EACH PARTICIPATNS AND ATTRIBUTE TYPES.....	197
FIGURE 7-8. INTERGRATION WINDOW OVERVIEW AND HALF-HEIGHT BANDWIDTH STATISTICS.....	198
FIGURE 8-1. CONCEPTUAL FRAMEWORK INTEGRATING ALL RESULT IN THE THESIS.	205
FIGURE 9-1. INDIVIDUAL RESULT CORRESPONDS TO THE FIRST ROW OF FIGURE 3-5.....	217
FIGURE 9-2. INDIVIDUAL RESULT CORRESPONDS TO THE SECOND ROW OF FIGURE 3-5.....	218
FIGURE 9-3. INDIVIDUAL RESULT CORRESPONDS TO THE THIRD ROW OF FIGURE 3-5.....	219
FIGURE 9-4. INDIVIDUAL RESULT CORRESPONDS TO THE FORTH ROW OF FIGURE 3-5.....	220
FIGURE 9-5. INDIVIDUAL RESULT CORRESPONDS TO THE FIRST ROW OF FIGURE 3-6.....	221
FIGURE 9-6. INDIVIDUAL RESULT CORRESPONDS TO THE SECOND ROW OF FIGURE 3-6.....	222
FIGURE 9-7. INDIVIDUAL RESULT CORRESPONDS TO THE THIRD ROW OF FIGURE 3-6.....	223
FIGURE 9-8. INDIVIDUAL RESULT CORRESPONDS TO THE FORTH ROW OF FIGURE 3-6.....	224
FIGURE 9-9. SINGLE TRIAL OUTPUT OF THE NAÏVE MODEL FOR SIGNAL INTERVAL INPUT	225
FIGURE 9-10. SINGLE TRIAL OUTPUT OF THE NAÏVE MODEL FOR SIGNAL INTERVAL INPUT	226
FIGURE 9-11. INDIVIDUAL CCG FUNCTION FOR THE FIGURE 6-2.....	227
FIGURE 9-12. INDIVIDUAL YL's CCG FUNCTION CORRESPONDS TO THE FIGURE 6-5	228
FIGURE 9-13. INDIVIDUAL YJC's CCG FUNCTION CORRESPONDS TO THE FIGURE 6-5	229
FIGURE 9-14. INDIVIDUAL YT's CCG FUNCTION CORRESPONDS TO THE FIGURE 6-5.....	230
FIGURE 9-15. INDIVIDUAL ZTS's CCG FUNCTION CORRESPONDS TO THE FIGURE 6-5	231
FIGURE 9-16. INDIVIDUAL ZWH's CCG FUNCTION CORRESPONDS TO THE FIGURE 6-5	232

List of Tables

TABLE 1-1 SUMMARY OF CHAPTER–QUESTION MAPPING.....	35
TABLE 3-1. PROPORTION OF CHOOSING TARGET OR INVERSE RESPONSE IN MIXED-WAVE CONDITIONS, WITHIN- ATTRIBUTE	80
TABLE 6-1. PEAK LATENCY STATISTICS BETWEEN PT AND AT ACROSS ATTRIBUTE TYPES.....	157
TABLE 7-1. PEAK LATENCY (IN MILLISECONDS) ACROSS DIFFERENT ATTRIBUTE TYPES AND TRACKING CONDITIONS.	191

List of Publication

Chapter 3

Chen, Y.J., Sun, Z., & Nishida, S. (2024). Feature-invariant processing of spatial segregation based on temporal asynchrony. *Journal of Vision*, 24(5), 15.
<https://doi.org/10.1167/jov.24.5.15>

Chen, Y.J., Sun, Z., Nishida, S. (2024). [poster]. Square versus triangular wave modulation in temporal asynchrony segmentation: an insight into underlying computation. Asia Pacific Conference on Vision, Singapore.

Chen, Y. J., & Nishida, S. (2023). [poster]. Temporal limits of visual segmentation based on temporal asynchrony in luminance, color, motion direction, and their mixtures. Annual meeting of vision science society 2023. Florida, USA.
<https://doi.org/10.1167/jov.23.9.4864>

Chapter 5

Chen, Y. J., Sun, Z., & Nishida, S. (in revision). Human visual grouping based on within-and cross-area temporal correlations. *PLOS computational biology*.
<https://doi.org/10.1101/2025.03.27.645664>

Chen, Y.J., Sun, Z., Nishida, S. (2024). [poster]. The Role of Cross-Area and Within-Area Temporal Correlations in Visual Segmentation. European Conference on Visual Perception 2024. Aberdeen, UK.

Chapter 6

Chen, Y.J., Sun, Z., & Nishida, S. (2025). Revealing temporal dynamics of the visuo-motor system via continuous tracking of position and attribute. *Journal of Vision*, 25(8), 19. <https://doi.org/10.1167/jov.25.8.19>

Chen, Y.J., Sun, Z., Nishida, S. (2024). [poster]. Temporal dynamics gap between position tracking and attribute tracking. Annual meeting of vision science society 2023. Florida, USA. <https://doi.org/10.1167/jov.24.10.465>

Chapter 7

Chen, Y.J., Sun, Z., Nishida, S. (2025). [poster]. Temporal dynamics of spatial segmentation based on temporal correlation measured by continuous tracking. Annual meeting of vision science society 2025. Florida, USA.

Others

Sun, Z., **Chen, Y.J.,** & Yang, Y.H., Nishida, S. (in press). Machine Learning Modeling for Multi-order Human Visual Motion Processing. Nature Machine Intelligence. <https://arxiv.org/abs/2501.12810>

Sun, Z., **Chen, Y.J.,** Yang, Y.H., Nishida, S. (2023) Modeling Human Visual Motion Processing with Trainable Motion Energy Sensing and a Self-attention Network. 2023 Conference on Neural Information Processing Systems, New Orleans, United States. <https://doi.org/10.48550/arXiv.2305.09156>

Sun, Z., **Chen, Y.J.,** Yang, Y.H., Nishida, S. (2023) Comparative Analysis of Visual Motion Perception: Computer Vision Models versus Human Vision. Conference on Cognitive Computational Neuroscience, Oxford, UK. https://2023.ccneuro.org/view_paper32e1.html?PaperNum=1080

Sun, Z., **Chen, Y.J.,** Yang, Y.H., Nishida, S. (2024). [oral]. Acquisition of second-order motion perception by learning to recognize the motion of objects made by non-diffusive materials. Annual meeting of vision science society 2023. Florida, USA. <https://doi.org/10.1167/jov.23.9.4894>

Sun, Z., **Chen, Y. J.,** Yang, Y. H., & Nishida, S. (2023). [poster]. Modeling of Human Motion Perception Mechanism: A Simulation based on Deep Neural Network and Attention Transformer. Annual meeting of vision science society 2023. Florida, USA. <https://doi.org/10.1167/jov.23.9.4894>

Chapter 1. General Introduction

We live immersed in a world of ceaseless variation—a continuous flux of light, color, and motion that streams across our visual field. From moment to moment, the environment shifts in form and intensity, in structure and rhythm. Yet, our perception of this world is not chaotic. It is, instead, remarkably coherent: objects persist, surfaces remain stable, and the world appears meaningful. This contrast—between the instability of external signals and the stability of internal experience—lies at the heart of visual perception. The visual system, though bounded by physiological limitations in spatial and temporal resolution, achieves an extraordinary feat: it transforms unstructured, noisy input into organized perceptual representations. It is not merely a passive transducer of retinal signals, but an active constructor of visual structure. This constructive role depends fundamentally on processes that extract and impose order on sensory input—processes that group, separate, align, and segment. Through such mechanisms, the visual system assigns elements to forms, defines boundaries between regions, and gives rise to the experience of coherent scenes. Somewhere in the processing hierarchy, the disordered stream of input gives way to structure. Understanding how such organization arises—how structure emerges from signal—forms the foundation of perceptual theory.

In this thesis, we focus specifically on the problems of perceptual grouping and segmentation, a subset of perceptual organization that concerns how elements in the visual field are partitioned into coherent units. While these processes are deeply intertwined with functional roles (Marr’s first level) and neural implementation (Marr’s third level), the present work adopts a computational perspective, referring specifically to the representational and algorithmic levels in Marr’s three-level framework—focusing on how visual signals are represented, transformed, and utilized to support task performance during processing (Marr, 2010). The functional and neurophysiological aspects are addressed separately in 0. Here, we begin by tracing the development of computational approaches to grouping and segmentation, from early theoretical proposals to more recent algorithmic models.

1.1. Computational Approach of Perceptual Grouping/Segmentation

The computational approach to visual grouping and segmentation can be traced back to the emergence of Gestalt psychology. A comprehensive historical account is provided by Wagemans et al. (2012) and further elaborated in the opening chapter of their edited volume (Wagemans, 2015), which covers a wide range of topics in mid-level vision. In the present section, however, we focus specifically on the subset of ideas most relevant to perceptual grouping and segmentation.

Gestalt psychology offers several general “principle” of guiding how the local features can be grouped together (Köhler, 1930; Wertheimer, 1923). For example,

- **Proximity** suggests that elements located close to each other in space are more likely to be grouped together than those farther apart.
- **Similarity** indicates that elements sharing visual attributes such as color, shape, size, or orientation tend to be grouped.
- **Good continuation** posits that elements aligned along a smooth curve or path are perceived as part of the same object.
- **Prägnanz** suggests that visual elements are grouped to form closed or complete figures, even when portions of the contour are missing.
- **Common fate** extends grouping into the temporal domain, proposing that elements moving in the same direction or changing together are perceived as a group.

Additional principles such as symmetry, common region, and connectedness further elaborate the conditions under which grouping is facilitated. These principles were formulated based on phenomenological observations—describing what is perceived rather than how perception is computed. They have been remarkably successful in accounting for a wide range of perceptual phenomena, including illusory contours and texture-based segmentation. Gestalt theory emphasizes that the visual system imposes structure and coherence on fragmented sensory input, organizing the environment into discrete, meaningful units. Nevertheless, despite their descriptive power and enduring influence, Gestalt principles provide limited insight into the algorithmic or neural mechanisms underlying perceptual organization (Jäkel et al., 2016). This gap has motivated the development of computational and representational models that seek to formalize these phenomena in a more operationally defined framework.

The first quantitative approach to perceptual grouping and segmentation is rooted in the study of texture segmentation. Texture segmentation concerns how the visual system performs a kind of statistical test across a presumed boundary. Such boundaries are typically categorized as first-order, second-order, or higher-order, depending on the local features involved. A first-order boundary is defined by a difference in the mean tendency of local features—analogueous to the first moment of a statistical distribution—such as luminance (Julesz, 1962, 1965), line orientation (Beck, 1966; Julesz, 1965; H. C. Nothdurft, 1985), though orientation is sometimes considered a second-order feature, as it requires sampling across at least two spatial positions. For example, a boundary defined by a mean luminance difference is illustrated in the left panel of Figure 1-1. In contrast, a second-order boundary refers to a boundary defined by variations in local features, corresponding to differences in the second moment (variance) of the distribution, while the mean remains similar. Such boundaries are also perceptible across a variety of features, including luminance (Dakin, 2015) and orientation (Dakin & Watt, 1997; Motoyoshi & Nishida, 2001, 2002), as shown in the right panel of Figure 1-1. Meanwhile, the second-order boundary can also be referred to as the boundary based on “non-simple” features such as the V, T, L, and X junction such kind of corner-like features (Beck, 1966). The higher-order statistics have been implicated in other perceptual tasks, such as material perception (Motoyoshi et al., 2007), while texture segmentation is primarily based on first- and second-order boundaries (Dakin, 2015). Building upon this statistical framework, psychometric functions that consider both the mean and variance of feature distributions have been shown to closely replicate human sensitivity to texture differences, particularly for orientation-based segmentation (Dakin, 2001; Dakin & Watt, 1997). The underlying idea assumes that the visual system samples a set of points across the visual field and evaluates whether the features from two hypothesized regions are statistically different. Segmentation is thus conceptualized as a decision process based on whether the observed difference in feature distributions across the boundary is sufficiently large to warrant perceptual separation (Rosenholtz, 2000).

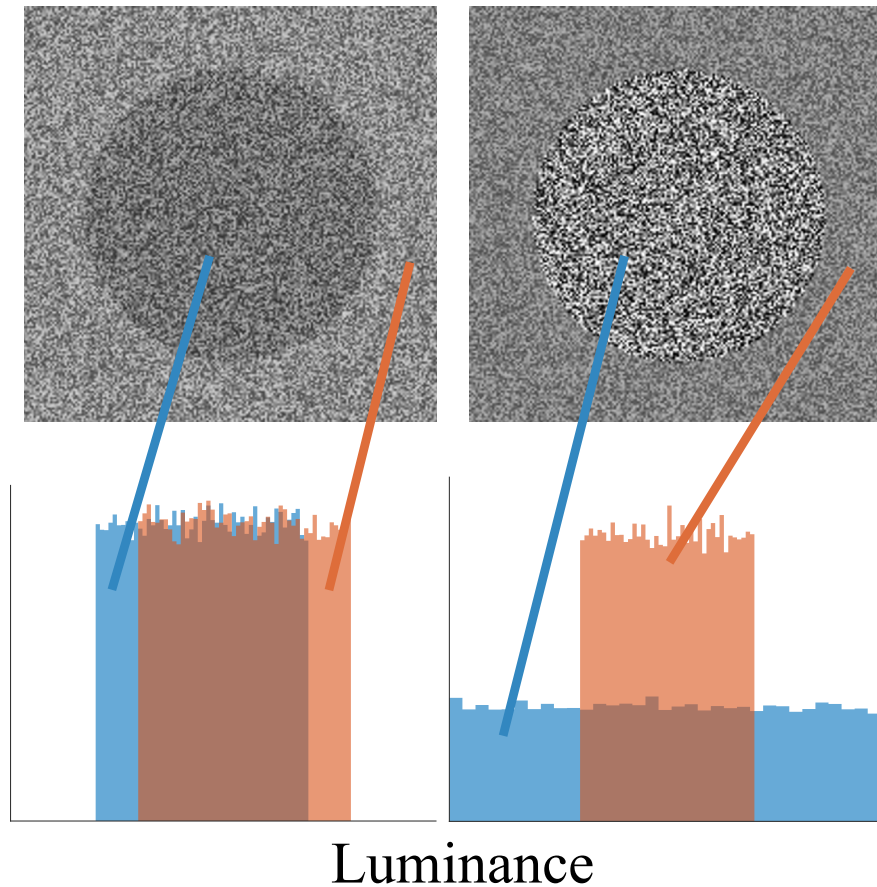


Figure 1-1. Demonstration of 1st and 2nd order luminance statistics based segmentation. The left panel illustrates a segmentation based on a first-order luminance difference, where the mean luminance of the target region differs from the background. The right panel shows segmentation based on a second-order luminance difference, where the mean luminance remains similar across regions, but variance differs. Histograms below each image depict the corresponding luminance distributions for the foreground (blue) and background (orange) regions, highlighting the statistical basis of the segmentation.

The second major quantitative approach to perceptual grouping and segmentation arises from the tradition of image-processing models. Image-processing models attempt to take the original image as input and, through a sequence of relatively simple operations, replicate the perceptual experience and behavior associated with visual segmentation (J. R. Bergen & Adelson, 1988; James R. Bergen & Landy, 1991; Kingdom et al., 2003; Michael S. Landy & Graham, 2003; Malik & Perona, 1990; Rosenholtz, 2000). The most prominent framework within this tradition is the so-called Filter–Rectify–Filter (FRF) model, also known as the Linear–Nonlinear–Linear (LNL) framework. In the FRF architecture:

- The first filtering stage acts as a simple feature extractor, mimicking the function of biological structures such as the center-surround receptive

fields in the lateral geniculate nucleus (LGN) or the orientation-selective line detectors in the primary visual cortex (V1).

- The rectification stage applies a nonlinear operator, usually a full-wave rectification (e.g., absolute value function, quadratic function, or other even functions), energy computation (Adelson & Bergen, 1985), or contrast normalization (Carandini & Heeger, 2012; Heeger, 1992).
- The second filtering stage acts as a coarse-scale feature detector, designed to detect broader boundaries or structures across the visual field by pooling rectified outputs.

This framework has been shown to successfully account for the detection of both first-order boundaries—where segmentation is based on mean feature differences, such as changes in luminance or orientation (Malik & Perona, 1990)—and second-order boundaries—where more complicated junction and terminal features are used (Bergen & Adelson, 1988; Dakin & Watt, 1997), or where segmentation arises from differences in the variance or energy of features across regions without a change in the mean (Graham & Sutter, 2000; Motoyoshi & Nishida, 2004; Wilson, 1993).

1.2. Primary Limitation of Current Computational Approach

Despite the explanatory power of the approaches described thus far, several fundamental issues in perceptual segmentation remain unresolved.

First, statistical-test models typically assume a candidate boundary and then compare the statistical distributions of visual features across hypothetical regions to determine whether segmentation is justified. While this statistical-comparison architecture plausibly reflects perceptual and decisional uncertainty during segmentation tasks, these models operate under the assumption of a given boundary and assess whether statistical differences justify its presence, but they offer limited insight into how such candidate boundaries are generated or selected in the first place. This limitation arises primarily because computing distribution differences inevitably collapses local shape and global structural information into a single summary statistic, thereby losing critical spatial details necessary for understanding boundary inference and region assignment.

Second, although boundary detection can be effectively achieved by FRF-type models, these models remain predominantly sensitive to local feature modulations and

thus have inherent difficulty integrating spatial information over larger image regions. Numerous psychophysical studies have demonstrated that long-range spatial integration, often spanning visual extents greater than 2 degrees, is critical for robust perceptual segmentation and grouping (Meese, 2010; Motoyoshi & Nishida, 2002; Wolfson & Landy, 1999). Such large-scale integration likely requires additional computational mechanisms, including multi-scale processing and fusion (James R. Bergen & Landy, 1991; Portilla & Simoncelli, 2000; Rodrigues & du Buf, 2006) or gradual evidence accumulation across spatial extents and temporal durations (Jeurissen et al., 2016). Furthermore, important structural cues encoded in the phase spectrum of visual stimuli, which go beyond conventional first- and second-order statistical features, have also been shown to significantly influence segmentation performance (Arsenault et al., 2011). Yet current implementations of the FRF framework do not effectively capture these phase-based structural cues, suggesting that additional mechanisms may be required for such integration.

Third, most existing computational models of segmentation and grouping have primarily focused on static visual cues, such as luminance, orientation, junctions, and spatial frequency differences. Consequently, we have limited understanding of how these models generalize when visual stimuli vary dynamically over time. It remains unclear whether existing frameworks can adequately describe human perceptual performance under conditions involving temporal variation or motion-defined structures. Given that dynamic stimuli represent arguably more ecologically relevant scenarios encountered in everyday vision, understanding the mechanisms underlying temporal perceptual grouping and segmentation becomes crucial.

In the following section, we thus shift our focus toward perceptual grouping and segmentation processes that are guided by dynamic visual cues, aiming to address these critical theoretical and empirical gaps.

1.3. Dynamic Cue for Perceptual Grouping/Segmentation

While traditional models of perceptual grouping and segmentation have primarily emphasized spatial configurations and static visual features, real-world visual experience typically unfolds dynamically over time. Visual scenes in daily life are rarely static. Instead, objects, surfaces, and boundaries are often defined and

disambiguated through their temporal properties, such as motion coherence, temporal synchrony, or correlated changes in visual attributes. In this section, we examine how dynamic features influence visual segmentation and evaluate whether existing models, which were originally developed for static inputs, can also explain human performance in tasks involving temporal synchrony.

One of the classical perceptual grouping phenomena based on dynamic cues is the principle of common fate, which states that visual elements moving in the same direction and at the same speed tend to be grouped together perceptually (Wertheimer, 1923). The visual system demonstrates a remarkable capacity to extract coherent sets of dots, patches, or objects, even under variations in direction, speed, and exposure duration (Sekuler, 1990; Uttal et al., 2000). This grouping ability is not limited to two-dimensional spatial layouts but also extends into three-dimensional configurations (Andersen & Bradley, 1998; Murray, 2003). While the original notion of common fate was primarily grounded in the coherence of motion signals, subsequent research has proposed a more general form of this principle, emphasizing temporal structure beyond physical displacement. For instance, Sekuler & Bennett (2001) introduced a checkerboard stimulus composed of 10-by-10 square patches, each initialized with randomized luminance values. Across time, the luminance of each patch fluctuated according to a sinusoidal function. Within this grid, a subset of six patches served as the target, defined by their luminance modulations being in phase with one another but 180 degrees out of phase with the background patches. The results showed that observers could reliably detect the target even when the luminance modulation amplitude was very low and such ability cannot be ascribed to the detection of the local luminance difference. Their study suggest that temporal coherence alone was sufficient to support perceptual grouping.

Perceptual grouping and segmentation based on temporal information (we abbreviated as temporal segmentation in the following) have been widely investigated since the 1990s (for a review, Blake & Lee, 2000; Farid, 2002). Early studies focused on the role of stimulus onset or offset timing in determining perceptual boundaries. For example, Ramachandran & Rogers-Ramachandran, (1991) presented displays in which targets and backgrounds had opposite luminance values (e.g., 0 vs. 1) that reversed on consecutive frames and were shown repeatedly. Their results demonstrated that observers could reliably perceive the boundary between target and background even at temporal reversal rates up to 17 Hz. Similarly, Fahle, (1993) used a 14-by-14 dot matrix

in which background dots were delayed relative to the target dots, and found that a temporal offset as small as 5 milliseconds was sufficient for segmentation. Although this type of segmentation based on onset asynchrony appears highly efficient, its effectiveness is strongly modulated by spatial frequency content. High spatial frequency stimuli are more easily segmented than low-frequency ones (Kojima, 1998). Later studies expanded this line of research to include other visual attributes, demonstrating that temporal offsets could support segmentation not only for luminance-defined elements (Forte et al., 1999; Usher & Donnelly, 1998), but also for stimuli defined by motion direction (Kandil & Fahle, 2001, 2004), orientation (Leonards et al., 1996; Leonards & Singer, 1998; Usher & Donnelly, 1998), and color (Leonards & Singer, 1998).

The segmentation phenomena described above can, in certain cases, be accounted for by the Filter-Rectify-Filter (FRF) framework introduced in the previous section. This is primarily because an onset asynchrony between the target and background inevitably introduces local differences in mean luminance and/or temporal contrast, allowing spatiotemporal filters to extract the resulting transients. For example, Figure 1-2. (A) illustrates a simplified version of the paradigm used by Fahle, (1993) where the target region exhibits a delayed luminance onset relative to the background. When processed through a simple spatiotemporal filter (middle panel), this temporal discontinuity produces a strong localized response (right panel), revealing the boundary. So as the general common fate generated by Sekuler & Bennett, (2001) (a demo referred to Figure 1-2. B). Several studies have argued that segmentation based on temporal cues—particularly onset or flicker differences—can be fully explained by such low-level spatiotemporal energy detectors (Adelson & Farid, 1999; Farid, 2002; Farid & Adelson, 2001; Morgan & Castet, 2002). Indeed, it is difficult to deny the involvement of early spatiotemporal filtering mechanisms in perceptual grouping and segmentation tasks involving dynamic input (Motoyoshi, 2004).

However, as spatial structure is progressively removed and only higher-order temporal relationships remain, the performance of the FRF model deteriorates significantly. Figure 1-2.(C) illustrates this limitation. In this condition, the stimulus is modified from the generalized common fate paradigm by randomizing the temporal contrast polarity across elements. Although the FRF framework still yields some intermediate activity, its final output fails to delineate the segmentation boundary clearly.

This contrasts with human perception, in which the boundary remains salient and easily detectable.

To succeed in this task, observers must judge whether adjacent elements change at the same time, irrespective of whether their luminance is increasing or decreasing. This type of temporal co-occurrence cannot be captured by simple spatiotemporal filtering. Supporting this view, S.-H. Lee & Blake, (1999) introduced the concept of temporal structure, a higher-order temporal statistic that reflects the synchrony of onset and offset events across multiple elements. Rather than relying on local signal energy, this mechanism evaluates global coherence in change timing. Subsequent work has emphasized the role of temporal structure in shaping perceptual segmentation (Blake & Lee, 2000; Guttman et al., 2005, 2007), and has shown that such information is not fully accessible to FRF models (Guttman et al., 2007), though occasional segmentation effects may still emerge under specific conditions (Adelson & Farid, 1999).

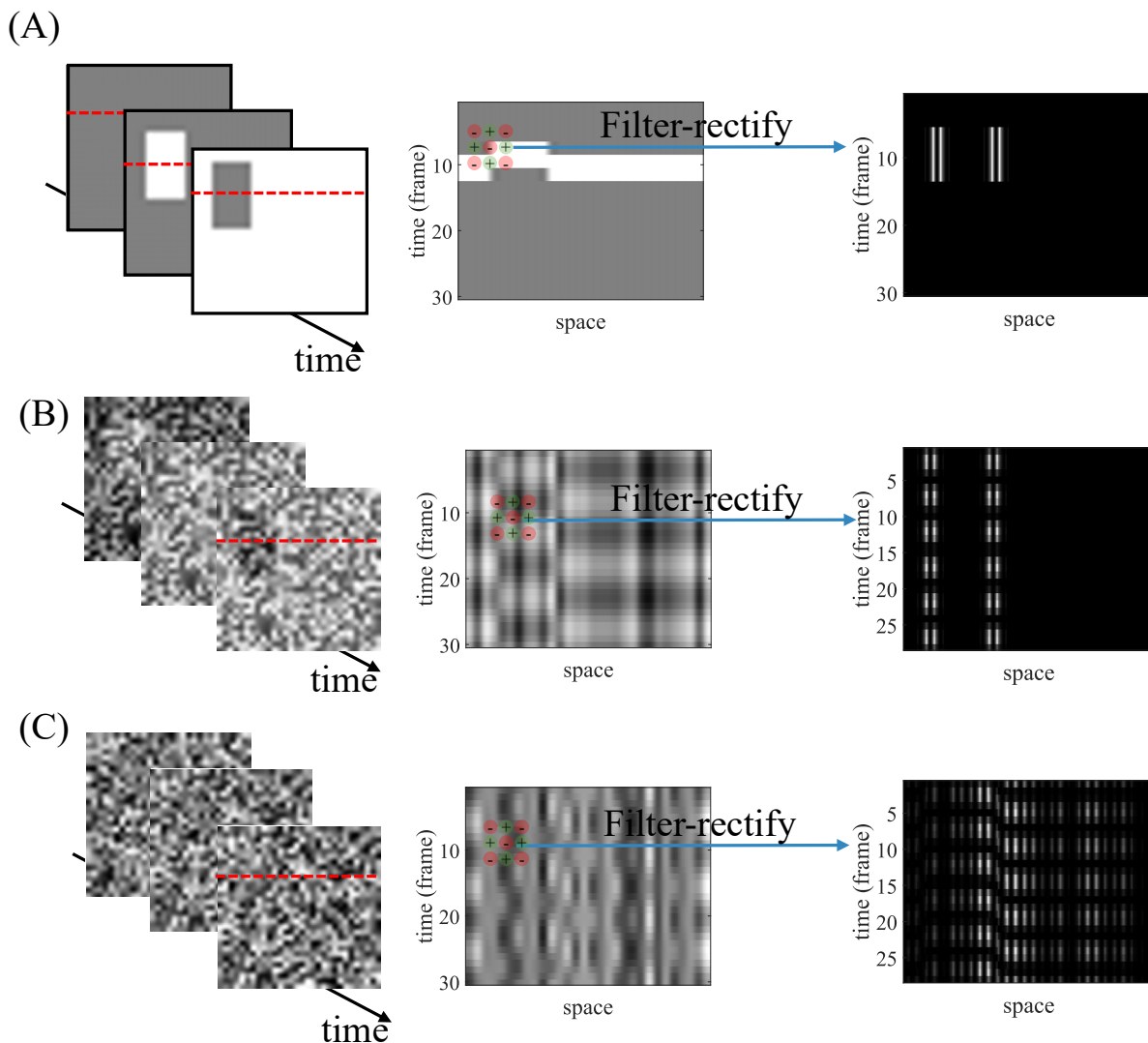


Figure 1-2. Demonstration of whether various spatiotemporal textures can be parsed using a simple Filter-Rectify (FRF) model.

(A) Simulated version of the stimulus structure used by Fahle, (1993), where segmentation relies on onset asynchrony between the target and background. The middle panel shows the corresponding space-time (X-T) representation, with the overlaid filter indicating the orientation of the spatiotemporal receptive field. The right panel shows the FRF model response, where clear boundary detection is possible. (B) Generalized common fate condition (Sekuler & Bennett, 2001), in which target vary synchronously in phase, while the background patches fluctuate with opposite phase. The FRF output shows detectable segmentation. (C) Similar to (B), but with the polarity of temporal variation (i.e., 0 or 180-degree phase) randomized across all patches. In this case, the FRF output fails to isolate a clear boundary.

Another challenge for existing models lies in accounting for the cross-attribute invariance and integration observed in temporal grouping and segmentation. Alais et al., (1998) presented a stimulus composed of four Gabor patches drifting back and forth, which could give rise to the percept of a global clockwise or counterclockwise rotation. Crucially, they manipulated whether the contrast of the Gabors flickered in synchrony or asynchronously. The results showed that Gabors with synchronous contrast flicker were more likely to be perceived as rotating globally. This finding demonstrates that

temporal synchrony across attributes, specifically, contrast and motion, can facilitate perceptual grouping beyond feature-specific channels. More generally, Guttman et al., (2005) extended this idea by presenting a 14-by-14 array of Gabor elements in which four visual attributes: contrast, spatial frequency, orientation, and spatial phase, were independently and randomly modulated. Each Gabor element changed its attribute value over time, following a designated temporal pattern. Importantly, the detection of a target region depended solely on differences in temporal structure, not on the specific attribute being modulated. Participants were able to reliably detect the target despite the random and heterogeneous feature content, indicating that the visual system utilizes an attribute-invariant computation that is sensitive to higher-order temporal regularities.

Nevertheless, it is widely recognized that the processing speed for different attributes is extensively heterogeneous. For example, we notice the temporal dynamics of chromatic stimuli is sluggish than achromatic one, including behavior (Burr & Morone, 1993; O'Donnell & Colombo, 2008; Uchikawa & Yoshizawa, 1993) and neurophysiological evidence (D'Souza et al., 2011; Mullen et al., 2010). Or the dynamic information (e.g. motion) and the static information is transmitted in different speed through M pathway versus P pathway, where the dynamic one is faster (Kaplan, 2003; Wörgötter et al., 1999). Therefore, it could be challenging how the temporal synchrony across the attribute is determined if the temporal course of the separated attribute is apparently different, they may not be matched to compared readily. As a result, Moutoussis & Zeki, (1997) showed that the physically aligned timing of the motion direction change and the color change is perceptually thought as asynchronous, in which the motion direction change is perceptually perceived delayed than the color change. Such phenomena demonstrate the difficulty of defining temporal synchronization, especially across different visual attributes.

To account for perceptual asynchrony, Nishida & Johnston, (2002) proposed that the observed delays between the perception of different visual attributes are not due to slower low-level processing or transmission delays in specific neural pathways—such as motion-selective neurons being inherently slower than those for color. Instead, they introduced the concept of a “temporal marker”, which refers to a higher-order temporal representation that defines the moment an event is registered by the visual system. According to this view, perceptual asynchrony reflects differences in when the system assigns temporal labels to distinct feature changes, rather than differences in when those

changes are first detected. In this framework, the perceptual delay of motion relative to color arises because the salient point at which a change in motion direction is marked occurs later than that for color. This temporal marker is then used to compare the timing of events across features or modalities, enabling judgments of simultaneity and temporal order.

Subsequent studies supported this idea by demonstrating that such temporal comparisons operate across a wide range of features—including color, luminance, and motion—as well as across sensory modalities (e.g., vision and audition) (Fujisaki & Nishida, 2010). Importantly, these comparisons exhibit a much coarser temporal resolution than feature-specific detection mechanisms. While low-level detectors can resolve temporal changes at frequencies exceeding 20 Hz, the resolution of cross-attribute temporal judgments is markedly lower, typically around 3–5 Hz. This has led to the proposal that the visual system operates with two parallel temporal mechanisms: a fast, attribute-specific system for detecting local changes, and a slower, attribute-invariant system for temporal alignment and comparison (Holcombe, 2009).

The observed distinction between fast, attribute-dependent detection and slow, attribute-invariant timing comparison offers a valuable framework for understanding temporal segmentation. As illustrated in

Figure 1-3, we hypothesize that temporal segmentation may operate via two functionally distinct layers: a rapid, low-level system that independently processes each visual attribute through spatiotemporal filtering and nonlinear transformation, and a slower, high-level system that integrates temporal information across different attributes. The former layer can account for basic segmentation phenomena such as onset asynchrony and motion coherence, where temporal structure is tightly coupled with spatial features, and the segmentation is readily perceived after the processing of the FRF framework, as we show in Figure 1-2 (A) and (B). However, more complex segmentation tasks—such as those requiring temporal comparison across distinct visual attributes or involving temporal boundaries without reliable spatial cues (e.g., Figure 1-2 (C))—appear to exceed the computational capacity of the low-level attribute-specific system. These cases likely depend on a slower, higher-level mechanism that is attribute-invariant and operates across channels. As illustrated in

Figure 1-3. Conceptual framework of two level of temporal segmentation., we hypothesize that this mechanism comprises three functional components: (1) a feature integrator that combines temporal signals across attribute channels; (2) a synchrony comparator that evaluates temporal coherence over extended durations; and (3) a grouping module that determines the global assignment of each spatial sample to a coherent perceptual unit. The present thesis focuses on characterizing this high-level temporal segmentation process and its behavioral and computational signatures.

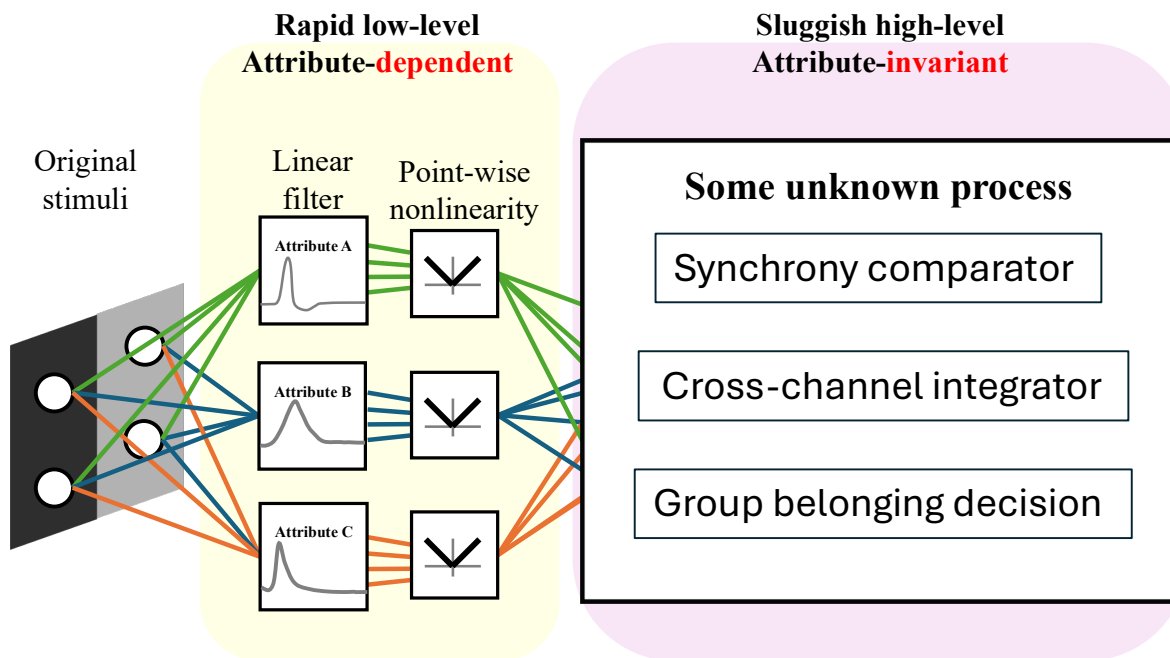


Figure 1-3. Conceptual framework of two level of temporal segmentation.

The schematic illustrates a conceptual framework for temporal segmentation that involves two distinct processing stages: a rapid, low-level attribute-dependent stage (yellow region) and a sluggish, high-level attribute-invariant stage (pink region). Original stimuli, composed of multiple spatial locations and visual attributes (A, B, C), are first processed through linear filters and point-wise nonlinearities, simulating feature-specific encoding. These outputs are then passed to a higher-level process whose internal computations remain unspecified but are hypothesized to include a synchrony comparator, a cross-channel integrator, and a group belonging decision module.

In this thesis, I aim to examine the existence and nature of a high-level, attribute-invariant mechanism underlying temporal segmentation. To this end, the following four questions structure the core inquiries of this work:

- 1) Existence: Does a temporal segmentation process exist that operates independently of specific visual attributes?
- 2) Integration: How do signals from different visual attributes interact in this process—does the system integrate, prioritize, or segregate them?

- 3) Computation: For such a temporal segmentation capability, what is the underlying computational mechanism responsible for detecting segmentation boundaries from purely temporal cues?
- 4) Dynamics: Can we experimentally estimate the temporal properties of this higher-level mechanism, such as its integration time or response latency, in a behaviorally measurable way?

These questions provide the foundation for the experimental chapters that follow. In the next section (1.4), I outline how each chapter in this dissertation is designed to address one or more of these theoretical issues.

1.4. Research Motivation and Scaffolding

This thesis aims to understand how the human visual system performs segmentation based on dynamic temporal information, with a particular focus on whether such segmentation relies on an attribute-invariant computation. In the preceding sections, we reviewed how traditional models—ranging from spatial statistics to image-computable filters—can partially explain segmentation under static conditions, but struggle to account for phenomena involving purely temporal or cross-attribute cues. The central goal of this thesis is to clarify the nature of temporal segmentation under these complex conditions, a mapping table of the core question for each chapter can be found in Table 1-1.

Chapter 3 addressed the first foundational question: Does the visual system support temporal segmentation in a manner that is invariant to specific visual attributes? To investigate this, we extended the approach of Guttman et al., (2005) by incorporating a broader range of visual features—including luminance, color, and motion direction—thereby enabling a more comprehensive assessment of attribute invariance. Temporal segmentation performance was quantified by measuring temporal frequency tuning curves for each attribute individually, as well as for combinations of attributes under cross-attribute conditions. The results reveal that segmentation performance across luminance, color, and motion direction exhibits similar frequency tuning (with peak sensitivity typically observed between 1–5 Hz), provided that the temporal transiency of the signal is preserved. Notably, this similarity was found in both within-attribute

and cross-attribute conditions, supporting the existence of an attribute-invariant mechanism underlying temporal segmentation.

Chapter 4 extends the investigation by examining how signals from different visual attributes interact during temporal segmentation. Specifically, we ask whether information from one feature domain can interfere with—or facilitate—segmentation based on another. Using the same class of stimuli introduced in later Chapter 5, we assessed segmentation performance as a function of temporal correlation, under conditions in which noise signals originated either from the same or from different feature dimensions. The results indicate that, when signal magnitude is equated across attributes, noise from different attribute channels produces interference comparable to that from within the same channel. A subsequent computational model demonstrates that simple linear summation across independent attribute channels accounts for the majority of the variance in human performance. These findings suggest that temporal segmentation may rely on an amodal integrative mechanism, which combines signals across attributes before computing temporal coherence for segmentation.

Chapter 5 builds upon the findings of the previous chapters by exploring the computational mechanisms that may underlie attribute-invariant temporal segmentation. To this end, we introduced stimuli generated using a novel Vision Transformer-based generator, specifically designed to minimize spatial cues and equate temporal frequency content across conditions—thereby controlling for low-level, attribute-dependent tuning effects. Within these stimuli, we systematically modulated the temporal correlation structure, hypothesized to serve as an attribute-invariant cue for representing temporal simultaneity during segmentation. Behavioral results indicated that participants’ performance closely tracked the magnitude of temporal correlation, regardless of the specific attribute. Motivated by this observation, we applied principles from graph spectral theory to develop a simple, image-computable model. Simulations based on this model closely matched human performance, suggesting that graph-based, attribute-invariant computations may plausibly underlie temporal segmentation in the human visual system.

Chapter 6 deviates from the central research questions but plays a critical role in addressing the final issue: the temporal dynamics underlying temporal segmentation. This chapter introduces a novel psychophysical paradigm—continuous tracking—in

which participants are asked to continuously track either the perceived location or the perceived content (e.g., attribute value) of a visual target. This method offers high sensitivity to temporal delays in perceptual processing. By comparing performance across different attributes and task demands, we sought to disentangle the underlying components contributing to system latency. The results indicate that tracking the location and tracking the content engage similar perceptual mechanisms, with the primary differences arising from the mapping between perception and motor output. Moreover, the latency estimates derived from continuous tracking were found to be highly consistent with those obtained via traditional reaction time measurements.

Finally, Chapter 7 revisits the question of temporal dynamics in temporal segmentation. Using the same stimulus generation algorithm introduced in Chapter 3, we first examined how reaction time varies as a function of temporal correlation strength. This served to test whether the visual system accumulates correlation strength as a cue for triggering perceptual decisions. To obtain a more fine-grained temporal profile, we further applied the continuous tracking method developed in the previous chapter. Three nested experimental conditions were designed to progressively isolate and estimate the effective temporal integration window for segmentation. An accompanying computational model was used to fit the tracking data and derive a pseudo-temporal window, which was then extended to explain the reaction time results. The findings demonstrate that the estimated temporal windows are consistent across different visual attributes, providing converging evidence for a shared temporal integration mechanism in attribute-invariant segmentation.

Table 1-1 Summary of Chapter–Question Mapping

Chapter	Research Focus	Corresponding question
Ch3.	Testing whether temporal segmentation operates independently of specific visual attributes	1) Existence
Ch4.	Investigating interactions between signals from different attributes during temporal segmentation	2) Interaction
Ch5.	Identifying the computation underlying attribute-invariant segmentation via a graph-based model	3) Computation
Ch6.	Introducing a continuous tracking method to decompose perceptual and action-related temporal processes	4) Dynamics
Ch7.	Estimating temporal integration windows using reaction time and tracking under different correlation levels	4) Dynamics

Chapter 2. Theoretical Background

To interpret visual grouping and segmentation from a computational perspective, it is essential to consider not only how such processes can be formally modeled, but also what functional roles they serve within the visual system and how they may be implemented in the neural architecture. This chapter begins with a brief review of the concept of mid-level vision and the neurophysiological basis of perceptual grouping and segmentation. These perspectives provide a broader context for the computational approach adopted in this thesis. Following this, we introduce the theoretical foundations necessary to understand the experimental paradigms and modeling strategies used in subsequent chapters. The overall aim is to formalize perceptual grouping as a structured computation, grounded in both low-level temporal encoding and high-level relational modeling.

Section 2.1 examines the functional role of perceptual organization within the visual hierarchy, emphasizing how grouping facilitates the interpretation of complex visual scenes by reducing ambiguity and supporting object-based representation.

Section 2.2 reviews neurophysiological evidence for grouping and segmentation, highlighting the diversity of cortical regions involved and the ways in which different attributes are encoded and integrated across the visual system.

Section 2.3 introduces the fundamental principles of temporal signal processing in vision, beginning with the framework of linear time-invariant (LTI) systems and extending to known physiological and psychophysical results across visual attributes.

Section 2.4 presents key concepts from graph theory, offering a formal framework to describe similarity-based grouping and spatial segmentation.

Section 2.5 outlines the methodology of continuous psychophysics, with a particular focus on cross-correlogram (CCG) analysis as a tool for estimating temporal system dynamics.

Together, these five sections establish a shared conceptual and analytical foundation for examining stimulus design, computational modeling, and human behavioral responses in the chapters that follow.

2.1. The Functional Role of Perceptual Organization in Vision Hierarchy

In conventional and practical terms, the visual processing hierarchy is commonly divided into three levels: low-level, high-level, and an intermediate stage known as mid-level vision (Anderson, 2020; Gilchrist, 2022; Nakayama et al., 1995), each characterized by distinct types of computation and representation. Low-level vision refers to simple processing in which single neurons respond selectively and locally to specific visual features—such as a line of a particular orientation—within a restricted retinotopic region. At this level, neural interactions are spatially limited, including mechanisms such as local association fields for contour integration (Field et al., 1993, 2013), lateral inhibition between retinal amacrine and bipolar cells (Cook & McReynolds, 1998; Yang & Wu, 1991), and center-surround antagonism in the lateral geniculate nucleus (Irvin et al., 1993; Nakayama, 1971; Westheimer, 2004).

In contrast, high-level vision operates over structured and semantically meaningful representations. It is concerned with the categorization and recognition of specific object classes—such as faces, scenes, and tools—and relies on regions that exhibit category-selective responses largely invariant to low-level features such as position, size, or viewpoint. For instance, the fusiform face area (FFA) is preferentially engaged during face perception (Kanwisher et al., 1997; SERGENT et al., 1992), while the parahippocampal place area (PPA) is implicated in processing visual scenes (Epstein & Kanwisher, 1998). These high-level representations are thought to depend critically on inferential mechanisms that associate current perceptual input with accumulated prior knowledge and memory, a view that can be traced back to the likelihood-based theory of perception proposed by Helmholtz (1872).

Compared to low- and high-level vision, mid-level vision is concerned with organizing local visual signals into perceptually coherent structures. These organizational computations serve as a functional bridge between early feature detection and high-level recognition, enabling the construction of coherent visual representations from fragmented input. Core functions of mid-level vision include grouping and segmentation (Roelfsema, 2006), border ownership assignment (Qiu et al., 2007), occlusion inference and contour interpolation (Heitger et al., 1998; Kellman & Shipley, 1991), surface construction (Nakayama et al., 1995), and long-range contour integration

(Jolicoeur et al., 1986), and even material perception (Anderson, 2020). For a comprehensive review, see Peirce (2015), Wagemans et al., (2012), or Kubilius et al., (2014).

Mid-level vision interacts bidirectionally with both lower and higher levels of processing, engaging in feedforward as well as feedback loops. On the one hand, low-level visual mechanisms can be modulated by mid-level computation. For example, colinear facilitation—in which the alignment of local orientation signals lowers the contrast threshold for contour detection—demonstrates how low-level mechanisms is supported via good continuation (Polat & Sagi, 1993). Similarly, evidence suggests that recurrent interactions between V1 and V2 are involved in boundary completion (Neumann & Sepp, 1999).

On the other hand, high-level processes can also modulate mid-level representations. For instance, object familiarity has been shown to facilitate the detection of shape and accelerate object recognition (S. Lee et al., 2010; Manahova et al., 2020; Peterson & Gibson, 1994). Moreover, the face inversion effect (Yin, 1969) illustrates how disrupting the expected high-level configuration of a stimulus can impair mid-level shape detection, suggesting top-down influences that extend beyond mere category recognition.

Taken together, these interactions suggest that mid-level vision functions as a perceptual hub, integrating low-level signals and preparing them for high-level interpretation (Rensink, 2000; Roelfsema, 2006). Once signals are integrated into an intermediate representation, referred to as the pre-semantic level (Anderson, 2020) or proto-object (Driver et al., 2001), this provides a stable, spatially coherent structure upon which functional operations such as attentional prioritization and visuomotor planning can reliably act. For example, it enables attentional prioritization (Driver et al., 2001; Egly et al., 1994, for a review, see Malcolm & Shomstein, 2015); and provides actionable targets for visuomotor interactions such as tracking, reaching, and grasping (Gaveau et al., 2014; Hu et al., 2022; Pisella et al., 2000). An analogous organizational role emerges in the temporal domain, where event segmentation partitions continuous experience into temporally bounded units that support memory and prediction (Zacks & Swallow, 2007).

In summary, perceptual organization plays a pivotal role in structuring visual input across all levels of the hierarchy. It provides the necessary intermediate representations

that support attention, action, and the mapping from sensory input to semantic knowledge. In this thesis, I focus on one of its most fundamental computational operations: grouping and segmentation. These processes serve as the entry point through which the visual system parses complex inputs into spatially and temporally coherent units.

2.2. Neural Basis for Perceptual Grouping/Segmentation

Neurophysiological evidence indicates that visual grouping and segmentation are subserved by a distributed network of cortical regions, with specific areas recruited depending on stimulus complexity and the nature of the visual feature involved. In general, these processes engage a set of hierarchically organized areas, including the primary visual cortex (V1), early extrastriate regions (V2–V4), the lateral occipital cortex (LO), the intraparietal sulcus (IPS), the medial temporal area (MT/V5), and the inferior temporal cortex (IT). Segmentation-related activity is frequently characterized by recurrent interactions across these areas, reflecting the dynamic and integrative nature of perceptual organization (Lamme & Roelfsema, 2000; Roelfsema et al., 2002). A simple location of each area can be referred to .

A general introduction to these specific brain areas follows the progression from the processing of simple to more complex visual features. However, it is important to note that the feature-specific responses discussed here are not solely the result of serial, feed-forward processing. Instead, they emerge from dynamic interactions between feedforward and feedback signals, and the precise neural mechanisms underlying purely feature-specific processing remain an active area of research. The primary visual cortex (V1) is the earliest extensively studied visual area and has been shown to support simple pattern detection, particularly based on two-dimensional frequency components, including the orientation and thickness of lines (Hubel & Wiesel, 1962). Complex cells in V1 are also responsive to features such as corners and junctions. Moreover, V1 receives feedback from several higher visual areas. Feedback from V2 is implicated in modulating local receptive fields to support contour integration and boundary detection (Neumann & Sepp, 1999), while feedback from V3 is associated with filling-in processes within bounded regions (Mendola et al., 2006), and feedback from V4 contributes to figure-ground segmentation (Poort et al., 2012).

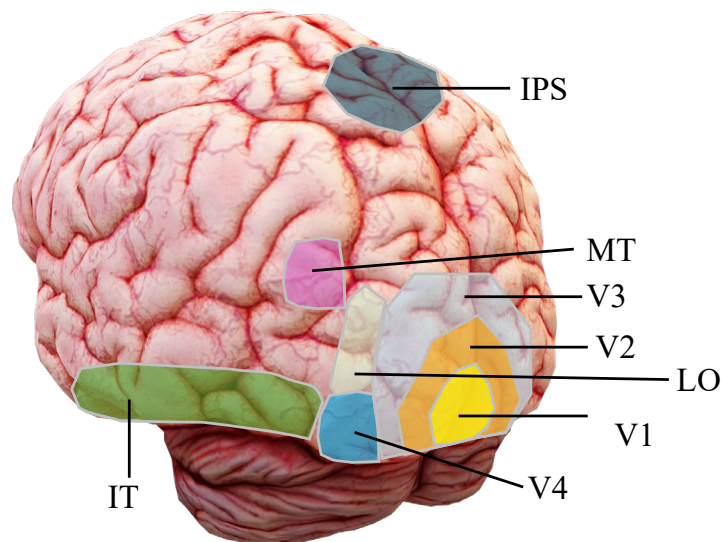


Figure 2-1. Cortical regions involved in perceptual grouping and segmentation.

Regions shown include the primary visual cortex (V1), secondary visual cortex (V2), tertiary visual cortex (V3), visual area V4, lateral occipital cortex (LO), middle temporal area (MT/V5), intraparietal sulcus (IPS), and inferior temporal cortex (IT). Anatomical locations are approximated based on typical retinotopic organization and functional imaging studies. The brain model is the built-in 3D model within Microsoft PowerPoint.

V2 lies immediately downstream of V1 in the visual hierarchy and receives strong feedforward input from it. It also receives binocular input and contains binocular-sensitive neurons, making it a key contributor to stereoscopic depth processing (von der Heydt et al., 2000). In particular, V2 has been implicated in the extraction of more complex texture structures, including second-order boundaries defined by contrast or other modulations rather than luminance (El-Shamayleh & Movshon, 2011; Song & Baker, 2007). Moreover, neurons in V2 exhibit selectivity for border ownership, indicating whether an edge belongs to a figure or background (H. Zhou et al., 2000), and are also responsive to illusory contours, such as Kanizsa figures (von der Heydt et al., 1984). In addition, V2 has been shown to be sensitive to the similarity structure across neighboring V1 units, suggesting a role in integrating local feature correlations into more global perceptual representations (Freeman et al., 2013).

V3 and the lateral occipital cortex (LO) occupy different levels of the visual hierarchy and belong to distinct processing streams but both contribute to the perception of spatially extended structures—V3 through early-stage contour integration and LO through higher-level shape coherence recognition. V3, located anterior to V2, is

responsive to extended contours and exhibits sensitivity to collinearity among spatially distributed edge elements, supporting the early stages of contour formation across considerable distances—even across the entire visual field (Li et al., 2019; Schira et al., 2004; Wu et al., 2023). In parallel, LO—a higher-tier region in the ventral visual stream—shows strong responses to coherent, object-like configurations, and is particularly tuned to grouped elements that form closed contours or exhibit symmetry (Kourtzi & Kanwisher, 2001; Volberg & Greenlee, 2014).

V4 and the inferior temporal cortex (IT) contribute to the representation of complex shape features and the abstraction of visual forms across transformations. V4, located downstream of V3 in the ventral visual pathway, exhibits selectivity for curvature and contour complexity (Pasupathy & Connor, 2001, 2002), and is also sensitive to three-dimensional surface structure inferred from shading and texture gradients (Yamane et al., 2008). In addition, neurons in V4 respond to inferred contours of occluded shapes, suggesting a role in surface-based completion (Bushnell et al., 2011). Feedback from V4 to V2 has also been proposed to support larger-scale border ownership assignment, further reinforcing the recursive nature of shape segmentation (Craft et al., 2007). Further along the hierarchy, IT plays a central role in encoding basic object shape categories. Neurons in IT exhibit tolerance to variations in size, position, and viewpoint, suggesting that they represent shape-based identity rather than raw feature conjunctions (Kourtzi & Kanwisher, 2001; Op de Beeck et al., 2001; Tanaka, 1996). In some cases, IT neurons also show category-selective responses to specific object classes, such as faces, places, or tools, as discussed in previous sections.

Finally, the middle temporal area (MT or V5) and the intraparietal sulcus (IPS) contribute to perceptual organization by processing dynamic and spatially structured information. MT is a motion-selective region that plays a central role in encoding the velocity, direction, and coherence of motion signals (Newsome et al., 1995; Orban et al., 1995). It supports not only the detection of moving stimuli but also the integration of local motion vectors into coherent global motion patterns, which is essential for segmenting objects defined by structure-from-motion or dynamic boundaries (Born & Bradley, 2005; Murray, 2003). IPS, a higher-order region in the dorsal stream, has been implicated in attention-mediated integration of multiple visual features and in representing hierarchical spatial relationships within complex visual scenes (Konen & Kastner, 2008; Xu & Chun, 2007). This area is thought to coordinate object-level

segmentation by combining information from both motion and feature-based pathways, especially under task demands that require selective attention or flexible grouping.

In summary, the neural basis of perceptual segmentation reflects a hierarchical architecture in which progressively more complex feature components are extracted across visual areas. These components are processed in a distributed manner, with distinct cortical regions specializing in different aspects of segmentation—from low-level edge detection to high-level shape and motion integration. Crucially, the integration of these features into coherent perceptual units appears to depend on attention-related mechanisms, which modulate and coordinate activity across these regions to support object-based organization.

Having examined the neural implementation of perceptual grouping and segmentation, the next step is to understand the computational algorithms that the visual system might employ. In other words, how exactly does the brain translate neural signals and intermediate-level representations into explicit segmentation decisions? To address this computational question, we turn to quantitative approaches derived from Gestalt principles and image-processing theories, which explicitly formalize grouping operations and segmentation outcomes.

2.3. Temporal Processing in The Visual System

To understand how the visual system processes time-varying input, it is essential to characterize its temporal dynamics. These dynamics determine how visual signals are temporally filtered, delayed, and integrated, and thus directly shape the system's output. A common approach is to model the visual system as a linear time-invariant (LTI) system, allowing researchers to characterize its behavior through either a temporal impulse response function (IRF) or its frequency transfer function, which captures how the system responds to transient inputs across time (Watson, 1986).

2.3.1. Linear time invariant system approach

A system is classified as linear and time-invariant (LTI) if and only if it satisfies the following two properties:

- **Linearity:** The system obeys both homogeneity (scaling) and superposition (additivity). Formally, for any inputs $x_1(t)$ and $x_2(t)$, and scalars a and b , the system $\mathcal{L}$ satisfies:

$$\mathcal{L}[ax_1(t) + bx_2(t)] = a\mathcal{L}[x_1(t)] + b\mathcal{L}[x_2(t)] \quad (2-1)$$

- **Time invariance:** The system's behavior does not change over time. That is, if input $x(t)$ yields output $y(t)$, then a time-shifted input $x(t - \tau)$ produces a time-shifted output $y(t - \tau)$ for any shift τ :

$$\forall \tau \in \mathbb{R}, \quad \mathcal{L}[x(t - \tau)] = y(t - \tau) \quad (2-2)$$

In many applications in vision science, a third assumption is often included:

- **Causality:** The system does not respond to future inputs. In other words, the output at any time t depends only on values of the input up to time t , not beyond. Formally:

$$y(t) = \mathcal{L}[x(\tau)], \quad \text{where } \tau \leq t \quad (2-3)$$

These assumptions simplify the analysis of temporal dynamics and provide a basis for estimating the system's impulse response and corresponding frequency-domain behavior, as discussed below.

2.3.1.1. Impulse response and convolution

In LTI systems, the most fundamental characterization is given by the impulse response, denoted as $h(t)$, as defined as follows:

$$h(t) = \mathcal{L}[\delta(t)] \quad (2-4)$$

which describes how the system responds to a unit impulse $\delta(t)$, a function with all its energy concentrated at $t = 0$. To generalize this to arbitrary input signals, note that any signal $x(t)$ can be expressed as a continuous weighted sum of shifted impulses:

$$x(t) = \int_{\tau=-\infty}^{\infty} x(\tau) \cdot \delta(t - \tau) d\tau \quad (2-5)$$

Applying linearity:

$$\mathcal{L}[x(t)] = \mathcal{L}\left[\int_{\tau=-\infty}^{\infty} x(\tau) \cdot \delta(t - \tau) d\tau\right] = \int_{\tau=-\infty}^{\infty} x(\tau) \cdot \mathcal{L}[\delta(t - \tau)] d\tau \quad (2-6)$$

Using the time-invariance property, we know that $\mathcal{L}[\delta(t - \tau)] = h(t - \tau)$, leading to the convolution integral:

$$\mathcal{L}[x(t)] = y(t) = \int_{\tau=-\infty}^{\infty} x(t) \cdot h(t - \tau) d\tau = x(t) * h(t) \quad (2-7)$$

Thus, the output is fully described as the convolution of the input with the system's impulse response.

2.3.1.2. Eigenfunction property of LTI system

A powerful property of LTI systems is that complex exponentials of the form e^{st} , where $s \in \mathbb{C}$, serve as eigenfunctions of the system under convolution. This means that when an exponential function is used as the input to an LTI system $\mathcal{L}$, the output is the same exponential scaled by a complex constant:

$$\mathcal{L}[e^{st}] = a(s)e^{st}, a, s \in \mathbb{C} \quad (2-8)$$

This property is extremely powerful. It implies that each exponential component is preserved in form by the system, and only its magnitude and phase are modified. By evaluating the system's response to a family of exponentials with different values of s , one obtains a corresponding set of scalar responses $a(s)$, each uniquely associated with the input e^{st} . This process defines a mapping from the time domain to a complex frequency domain, forming the basis of the Laplace transform:

$$x(t) \rightarrow X(s) = \int_{-\infty}^{\infty} x(t)e^{-st} dt \quad (2-9)$$

This transformation enables a frequency-based analysis of time-domain behavior.

In most practical cases, particularly when analyzing bounded, stable, and periodic signals, we typically restrict s to purely imaginary values, $s = i2\pi f$, yielding the Fourier transform:

$$\mathcal{F}\{x(t)\}: x(t) \rightarrow X(f) = \int_{-\infty}^{\infty} x(t)e^{-i2\pi ft} dt \quad (2-10)$$

By Euler's formula,

$$e^{i2\pi ft} = \cos(2\pi ft) + i\sin(2\pi ft) \quad (2-11)$$

The Fourier transform thus represents the signal in terms of its sinusoidal components, revealing its spectral composition in both amplitude and phase.

2.3.1.3. Duality and Relationship Between Impulse Response and Transfer Function

The impulse response $h(t)$ and its frequency-domain counterpart, the transfer function $H(f)$, are two fundamental representations of an LTI system's behavior. These two forms are mathematically equivalent, uniquely related through the Fourier transform and its inverse. This one-to-one correspondence ensures that a given impulse response maps to a single transfer function, and vice versa:

$$\begin{aligned} H(f) &= \mathcal{F}\{h(t)\} = \int_{t=-\infty}^{\infty} h(t)e^{-i2\pi ft} dt \\ h(t) &= \mathcal{F}^{-1}\{H(f)\} = \int_{f=-\infty}^{\infty} H(f)e^{i2\pi ft} df \end{aligned} \quad (2-12)$$

In addition, Fourier transforms exhibit the important property that convolution in the time domain corresponds to multiplication in the frequency domain. That is, for a system output $y(t) = x(t) * h(t)$, the corresponding frequency-domain representation satisfies:

$$Y(f) = X(f) \cdot H(f) \quad (2-13)$$

This result, often referred to as the convolution theorem, can be formally derived as follows. Starting from the definition of the Fourier transform:

$$\begin{aligned} Y(f) &= \mathcal{F}\{y(t)\} = \int_t y(t)e^{-i2\pi ft} dt \\ &= \int_t \left(\int_{\tau} x(\tau)h(t - \tau)d\tau \right) e^{-i2\pi ft} dt \end{aligned} \quad (2-14)$$

Applying Fubini's theorem to exchange the order of integration:

$$= \int_{\tau} x(\tau) \left(\int_t h(t - \tau) e^{-i2\pi f t} dt \right) d\tau \quad (2-15)$$

Using a change of variables $u = t - \tau \Rightarrow t = u + \tau$, $du = dt$, the expression becomes:

$$\begin{aligned} &= \int_{\tau} x(\tau) \left(\int_u h(u) e^{-i2\pi f(u+\tau)} du \right) d\tau \\ &= \int_{\tau} x(\tau) e^{-i2\pi f \tau} \left(\int_u h(u) e^{-i2\pi f u} du \right) d\tau \\ &= \left(\int_{\tau} x(\tau) e^{-i2\pi f \tau} d\tau \right) \cdot \left(\int_u h(u) e^{-i2\pi f u} du \right) \\ &= \mathcal{F}(x(t)) \cdot \mathcal{F}(h(t)) = X(f) \cdot H(f) \end{aligned} \quad (2-16)$$

This derivation confirms that the Fourier transform turns convolution into multiplication, allowing us to analyze system dynamics more efficiently in the frequency domain. Given this duality, researchers can choose to estimate either the time-domain impulse response or the frequency-domain transfer function, depending on the experimental design or analytical goal. Both representations describe the same underlying system, just in different domains. In the literature, both approaches have been widely used to probe the temporal characteristics of perceptual systems.

2.3.2. Impulse response measured by classical method

In this section, we review classical methods that have been used to extract the temporal dynamics of the visual system, along with their findings across different visual attributes. These methods operate under the assumption that the visual system behaves as a linear time-invariant (LTI) system. For clarity, we divide the discussion into two domains: temporal-domain-based approach, and frequency-domain-based approach.

2.3.2.1. Temporal-domain-based approach

Focus on the impulse response, three widely used paradigms include single-pulse stimulation, the double-pulse technique, and reverse correlation. In the single-pulse stimulation paradigm, a brief flash of light is presented to the subject (human or animal), and the resulting neural response is measured, typically via electrophysiological

recordings. This method allows direct observation of response latency and the temporal profile of activation in early visual pathways. The double-pulse technique extends this approach by presenting two sequential pulses at varying temporal intervals. By measuring the threshold amplitude required for the subject to perceive a flicker between the two flashes, researchers infer the temporal resolution of the system. If the interval is too short, the two pulses are fused and perceived as one; at longer intervals, the flicker becomes detectable. This technique provides a psychophysical estimate of the system's temporal integration window. A third method, reverse correlation, involves presenting stochastic (e.g., white noise) stimuli and computing the cross-correlation between the stimulus and the system's measured output (e.g., spike train, perceptual report). This approach effectively recovers the system's impulse response function by leveraging stimulus–response covariance. A similar rationale will be discussed in section 2.5, where we introduce the use of continuous tracking to recover temporal filters.

Importantly, all three methods have revealed a consistent pattern in the temporal processing of different visual attributes—particularly between achromatic and chromatic pathways. Early studies demonstrated that rod photoreceptors respond more rapidly to light onset than cone photoreceptors (Adrian, 1946). This latency difference propagates into the lateral geniculate nucleus (LGN), where magnocellular neurons (driven primarily by rods) exhibit faster response latencies than parvocellular neurons (driven by cones) (Maunsell et al., 1999). Psychophysical measurements corroborate these physiological findings by using the double-pulse technique (Burr & Morrone, 1993; Shinomori & Werner, 2003; Uchikawa & Yoshizawa, 1993). Similarly, reverse correlation analyses have confirmed that the impulse response to chromatic stimuli is broader and more sluggish than that to luminance stimuli (Okazawa et al., 2018; Ringach & Shapley, 2004; Zele et al., 2017).

2.3.2.2. Frequency-domain-based approach

In psychophysics, frequency-domain-based approaches are more commonly employed than time-domain-based methods. One primary reason is that it is often easier to obtain reliable and well-sampled measurements across temporal frequencies than across precise time points. In a typical frequency-based paradigm, participants are presented with flickering stimuli at specific temporal frequencies, and the task is to determine the minimum contrast at which the flicker becomes just detectable. This

yields a temporal contrast sensitivity function (tCSF), which conceptually corresponds to the system's transfer function—describing the relative sensitivity of the visual system across different temporal frequencies. In some studies, particular features of the temporal contrast sensitivity function (tCSF) are extracted to characterize system performance. One such feature is the critical flicker frequency (CFF), defined as the highest temporal frequency at which a flickering stimulus can still be perceived as temporally modulated, assuming maximal contrast. Beyond this frequency, the flicker is perceived as a steady light, indicating a breakdown in the system's ability to resolve rapid temporal changes (Landis, 1954).

One of the most prominent and well-established findings revealed by the temporal contrast sensitivity function (tCSF) is the distinction between the transient and sustained visual channels. This dual-channel framework was identified early in psychophysical research (Breitmeyer & Ganz, 1977; Harris & Georgeson, 1986; Kuliowski & Tolhurst, 1973; Legge, 1978). The transient channel is characterized by a bandpass temporal profile, optimally responding to rapid, dynamic changes in the input signal. In contrast, the sustained channel exhibits a low-pass profile, showing greater sensitivity to slow, gradual variations over time. Interestingly, these two channels also exhibit a spatiotemporal trade-off: while the transient channel offers higher temporal resolution, it tends to have a low-pass spatial profile; the sustained channel, conversely, exhibits bandpass spatial sensitivity, but lower temporal resolution. This functional distinction has been further supported by neuroimaging evidence—for example, recent fMRI studies have demonstrated corresponding temporal tuning properties in the lateral geniculate nucleus (LGN) (Horiguchi et al., 2009). It is also worth noting that the overall shape and amplitude of the tCSF can be influenced by several experimental and stimulus-related factors, including mean luminance, eccentricity, and stimulus duration. A comprehensive discussion of these modulatory factors, along with modeling approaches, can be found Ashraf et al. (2024) and Mantiuk et al. (2022). In addition to the transient–sustained channel distinction, more recent neuroimaging work has revealed attribute-specific temporal tuning: chromatic stimuli tend to activate neural responses that peak at lower temporal frequencies, whereas achromatic stimuli are tuned to higher temporal frequencies (D'Souza et al., 2011; Mullen, 2019; Mullen et al., 2010). This pattern mirrors the findings from time-domain-based methods, such as those based on

impulse response measurements, further confirming the temporal asymmetry between chromatic and achromatic visual processing.

2.4. Graph Theory in Perceptual Organization

2.4.1. Algebraic graph basis of organization task

In graph-theoretic approaches to perceptual organization, a visual stimulus is represented as a weighted undirected graph $G = (V, E, W)$, where each node $v_i \in V$ corresponds to a visual element, and the edge weight $W_{ij} \in \mathbb{R}$ quantifies the perceptual similarity between nodes v_i and v_j . These weights form a symmetric similarity matrix $W \in \mathbb{R}^{n \times n}$, which encodes the relational structure among visual elements. Each node v_i is associated with a feature vector $x_i \in \mathbb{R}^d$, which may represent stimulus attributes such as luminance, color, or temporal modulation. Pairwise similarities W_{ij} are computed from these features using a distance or affinity function. Two widely used formulations are:

(1) Gaussian kernel of Euclidean distance

$$W_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma^2}\right) \quad (2-17)$$

(2) Cosine similarity (normalized dot product)

$$W_{ij} = \frac{x_i^\top x_j}{\|x_i\| \|x_j\|} \quad (2-18)$$

Given such a similarity graph, perceptual grouping can be formalized as an optimization problem: the goal is to find a node-wise assignment function $f: V \rightarrow \mathbb{R}$ that reflects the perceived structure (e.g., grouping or segmentation), and that minimizes a global loss function ($\mathcal{L}(\cdot)$) reflecting the relational coherence defined by W . Such a task can be formalized as follows:

$$f(V) \rightarrow \operatorname{argmin}(\mathcal{L}(f(V), W)) \quad (2-19)$$

A broad class of such loss functions is based on the graph Laplacian $L = D - W$, where D is the degree with diagonal entries $D_{ij} = \sum_j W_{ij}$. The Laplacian encodes the

extent to which a function f varies across the graph structure. The canonical loss associated with L is the Laplacian energy functional, given by

$$\mathcal{L}(f, W) = f^\top L f = \sum_i \sum_j W_{ij} (f_i - f_j)^2 \quad (2-20)$$

This expression penalizes differences in f between nodes that are strongly connected, i.e., with large W_{ij} , a small value of $\mathcal{L}(f)$ indicate that similar nodes are mapped to similar values in f , corresponding to coherent perceptual grouping. Conversely, larger values reflect a more fragmented or segmented structure, where strongly connected elements receive dissimilar assignments.

In this way, the Laplacian energy serves as a smoothness constraint over the function f , and perceptual organization is modeled as the process of minimizing this energy. This formalism underlies many spectral and graph-cut-based approaches to segmentation and grouping. The precise way in which the optimal function f is computed—such as through eigenvalue decomposition or thresholding—depends on the specific model.

2.4.2. Relevance to perceptual grouping models

In practical applications of perceptual grouping, each node in the graph corresponds to a spatial sample, typically a pixel or a localized region (e.g., a patch or a superpixel) within the visual field. Each node v_i carries an associated feature vector $x_i \in \mathbb{R}^d$, which represents its local visual content, such as luminance, color, orientation, or temporal modulation profile. The goal of perceptual organization is then to derive a node function $f: V \rightarrow \mathbb{R}$ that assigns a scalar or vectorial value to each node, such that the resulting configuration respects the relational structure defined by the similarity matrix W . The optimization is guided by a loss function $\mathcal{L}(f; W)$, which quantifies the degree of structural coherence or segmentation achieved by the function f . Different models operationalize this framework with different assumptions about the form of f and the definition of $\mathcal{L}$. In this section, two representative approaches are examined:

1. The normalized cut algorithm: first proposed by Jianbo Shi & Malik (2000) which uses spectral decomposition of the graph Laplacian to construct a node function that minimizes normalized dissimilarity between groups.

2. Attention-based deep neural networks, particularly Vision Transformers, which learn both f and an implicit similarity structure directly from data using gradient-based optimization. The relationship between perceptual grouping and self-attention mechanisms is pointed out by Mehrani & Tsotsos (2023).

2.4.2.1. Normalized cut and graph spectrum

As introduced in the previous section, one of the simplest forms of segmentation involves separating two regions in comparable size, such as figure versus ground or target versus background. In this binary case, it has been mathematically proven that the optimal solution for partitioning a weighted graph under an L2 loss formulation corresponds to the second smallest eigenvector (Fiedler vector) of the graph Laplacian. We briefly outline the core idea of this result below.

For any undirected graph ($G = \{V, E, W\}$) with symmetric, positive weights, the (unnormalized) graph Laplacian is defined as:

$$L := D - W \quad (2-21)$$

The graph Laplacian $L \in \mathbb{R}^{n \times n}$ satisfies the following properties:

- L is real and symmetric: $L = L^\top$
- L is positive semi-definite: $x^\top L x \geq 0, \forall x \in \mathbb{R}^n$
- L is diagonally dominant. Its diagonal entries are non-negative, and its off-diagonal entries are non-positive due to the non-negative edge weights.

By the Spectral theorem for real symmetric matrices, these conditions imply that:

- There exists a full set of real eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n \in \mathbb{R}$
- There exists an orthonormal set of eigenvectors $\{\phi_1, \phi_2, \dots, \phi_n\} \subseteq \mathbb{R}^n$, which form a basis of $\mathbb{R}^n$

Accordingly, we can rewrite the graph Laplacian L as

$$L = \sum_{i=1}^n \lambda_i \phi_i \phi_i^\top \quad (2-22)$$

With eigenvalues ordered such that $0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$. The eigenvector corresponding to $\lambda_1 = 0$ (i.e., ϕ_1) is the normalized constant vector: $\phi_1 = \frac{1}{\sqrt{n}} [1, 1, \dots, 1]^\top$.

The number of zero eigenvalues of L is equal to the number of connected components in the graph. The second eigenvector ϕ_2 , known as the Fiedler vector, provides a relaxed solution to the optimal graph bipartition problem and is widely used in spectral clustering, and the corresponding eigenvalue λ_2 serves as an indicator of the graph's connectivity: a smaller λ_2 implies higher connectivity.

Lemma: Spectral relaxation of binary labeling with Laplacian energy

Let the ground-truth partition of the graph consist of exactly two non-empty, disjoint subsets, $A, B \subset V$ with $A \cup B = V, A \cap B = \emptyset$. Define the target labeling vector $T \in \mathbb{R}^n$ as:

$$T_i = \begin{cases} +1, & v_i \in A \\ -1, & v_i \in B \end{cases} \quad (2-23)$$

Define the Laplacian energy as the loss function based on equation (2-20), $Loss(T_i) = T^\top L T$. Minimizing this energy under the binary constraint $T_i \in \{\pm 1\}$ is NP-hard. In practice, we apply a spectral relaxation by relaxing T to a continuous vector: $\mathcal{F} := \{f \in \mathbb{R}^n: \|f\|^2 = n\}$.

Then:

1. If the labeling vector is balanced, i.e., $f^\top \mathbf{1} = 0$, the spectral relaxation of the constrained minimization problem is:

$$\min_f f^\top L f \quad (2-24)$$

Has its minimum attained at the Fiedler vector ϕ_2 , the second eigenvector of L .

2. If the labeling vector is unbalanced, i.e., $|A| = \alpha|B|$, for $\alpha \neq 1$, the the optimal relaxed solution f^* lies in the span of ϕ_1 and ϕ_2 , i.e.,

$$f^* = a_1 \phi_1 + a_2 \phi_2 \quad (2-25)$$

Proof.

Let $\{\phi_i\}_{i=1}^n$ be an orthonormal basis of eigenvectors of Laplacian matrix L , with corresponding eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$. By spectral theorem, any vector $f \in \mathbb{R}^n$ can be written as:

$$f = \sum_{i=1}^n a_i \phi_i, \sum_{i=1}^n a_i^2 = 1 \quad (2-26)$$

Substituting this into the Laplacian quadratic form, we obtain

$$\begin{aligned} f^\top L f &= \left(\sum_{i=1}^n a_i \phi_i \right)^\top L \left(\sum_{j=1}^n a_j \phi_j \right) \\ &= \sum_{i=1}^n \sum_{j=1}^n a_i a_j \phi_i^\top L \phi_j \end{aligned} \quad (2-27)$$

According to the properties of eigenvectors, $L\phi = \lambda\phi$ show:

$$\phi_i^\top L \phi_j = \phi_i^\top (\lambda_j \phi_j) = \lambda_j \cdot \phi_i^\top \phi_j \quad (2-28)$$

Since ϕ is an orthonormal basis,

$$\begin{aligned} \phi_i^\top L \phi_j &= \lambda_j \cdot \phi_i^\top \phi_j = \lambda_j \delta_{ij} \\ \delta_{ij} &= \begin{cases} 0, i \neq j \\ 1, i = j \end{cases} \end{aligned} \quad (2-29)$$

$$f^\top L f = \sum_{i=1}^n \sum_{j=1}^n a_i a_j \lambda_j \delta_{ij} = \sum_{i=1}^n \lambda_i a_i^2 \quad (2-30)$$

Case 1. Balanced labeling

If $f^\top 1 = 0$, then $f \perp \phi_1$, since $\phi_1 = \frac{1}{\sqrt{n}} 1$ spans the null space of L . Let $t := \frac{T}{\|T\|}$, so that $\|t\| = 1$ and $t \in \text{span}\{\phi_2, \dots, \phi_n\}$. Then:

$$t^\top L t = \sum_{i=2}^n \lambda_i a_i^2 \geq \lambda_2 \sum_{i=2}^n a_i^2 = \lambda_2 \quad (2-31)$$

With equality if and only if $t = \pm \phi_2$. Therefore, under the balance constraint and spectral relaxation $\|f\| = 1, f \perp \phi_1$, the optimal direction minimizing Laplacian energy is

$$f^* = \phi_2 \quad (2-32)$$

Case 2. Unbalanced labelling $|A| = \alpha|B|$

Now consider the case where the partition is unbalanced. Let the mean of T

$$\mu_T = \frac{1}{n} \sum T_i = \frac{|A| - |B|}{|A| + |B|} = \frac{\alpha - 1}{\alpha + 1} \quad (2-33)$$

The projection of T onto ϕ_1 is

$$\langle T, \phi_1 \rangle = \frac{1}{\sqrt{n}} \sum T_i = \sqrt{n} \mu_T = \sqrt{n} \frac{\alpha - 1}{\alpha + 1} \quad (2-34)$$

Normalizing T as $t := \frac{T}{\|T\|}$, with $\|T\| = 1$, we get

$$a_1 = \langle t, \phi_1 \rangle = \frac{\sqrt{n} \cdot T}{\sqrt{n}} = \frac{\alpha - 1}{\alpha + 1} \quad (2-35)$$

Assuming T is well-approximated by the first two eigenvectors (i.e., higher components are negligible), we write:

$$t \approx a_1 \phi_1 + a_2 \phi_2 \quad (2-36)$$

Using unit-norm condition $\|t\| = 1$, we get:

$$a_2 = \sqrt{1 - a_1^2} = \sqrt{1 - \left(\frac{\alpha - 1}{\alpha + 1}\right)^2} = \frac{2\sqrt{\alpha}}{\alpha + 1} \quad (2-37)$$

Thus, the normalized labeling vector is approximated by:

$$t \approx \frac{\alpha - 1}{\alpha + 1} \phi_1 + \frac{2\sqrt{\alpha}}{\alpha + 1} \phi_2 \quad (2-38)$$

2.4.2.2. Attention-Based Deep Neural Networks

While spectral methods segment images by analyzing the global structure of a pre-defined similarity graph, attention-based architectures, such as the Vision Transformer (ViT), adopt a purely data-driven strategy by dynamically learning relationships directly from input data. These models achieve segmentation by computing pairwise interactions between local image regions, implicitly forming a content-dependent fully connected graph, whose edge weights are learned through attention.

We briefly outline the core structure of ViT-based segmentation models below, noting that specific architectural details can vary across implementations.

Let an input image be represented as a tensor $X \in \mathbb{R}^{H \times W \times C}$, where H, W denote spatial dimensions, and C is the number of channels. The image is partitioned uniformly

into non-overlapping patches of size $P \times P$, yielding $N = \frac{H \cdot W}{P^2}$ total patches. Each patch is flattened and embedded into a learnable feature vector of dimension D , resulting in an initial sequence representation:

$$X^{N \times (P^2 C)} \mapsto Z_0 = [z_1^{D \times 1}, z_2^{D \times 1}, \dots, z_N^{D \times 1}]^T \in \mathbb{R}^{N \times D} \quad (2-39)$$

To preserve spatial information, positional embeddings $E_{pos} \in \mathbb{R}^{N \times D}$ (either fixed or learnable) are added

$$\widehat{Z}_0 = Z_0 + E_{pos} \quad (2-40)$$

The core operation is multi-head self-attention, iteratively applied across L transformer layers. At each layer l , three projection matrices $W^Q, W^K, W^V \in \mathbb{R}^{D \times d}$ map the embeddings into query (Q), key (K), and value (V) spaces:

$$Q = \widehat{Z}_l W^Q, \quad K = \widehat{Z}_l W^K, \quad V = \widehat{Z}_l W^V \quad (2-41)$$

The attention matrix $A \in \mathbb{R}^{N \times N}$ is then computed as:

$$A_{ij} = \frac{\exp\left(\frac{q_i^\top k_j}{\sqrt{d}}\right)}{\sum_m \exp\left(\frac{q_i^\top k_m}{\sqrt{d}}\right)}, \text{ with } \sum_{j=1}^N A_{ij} = 1 \quad (2-42)$$

Unlike the graph Laplacian (which is symmetric and static), the attention matrix A is asymmetric and varies with input content. The self-attention mechanism updates each token by aggregating values across all tokens:

$$Z_{l+1} = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V \quad (2-43)$$

Conceptually, this process can be seen as message passing on a fully connected graph $G = \{V, E, A\}$, with nodes corresponding to image patches, and dynamically computed edge weights reflecting learned query-key similarities.

After L such iterations, the final embedding $Z_L \in \mathbb{R}^{N \times D}$ encodes a global context-aware representation of each patch. For segmentation, this representation is reshaped to form a spatial feature map of size $\frac{H}{P} \times \frac{W}{P} \times D$. A lightweight decoder (e.g., bilinear upsampling or transposed convolution) reconstructs pixel-level label predictions from this map.

Let the predicted label distribution over K classes be denoted as $\hat{Y} \in \mathbb{R}^{H \times W \times K}$. The segmentation output at spatial location (i, j) is given by:

$$\widehat{Y}_{(i,j)} = \text{softmax}(\text{Decoder}(Z_L))_{(i,j)} \quad (2-44)$$

In summary, ViT-based segmentation models implicitly learn a dynamic, content-adaptive graph representation, where each image patch is a node, and relationships are attention-derived. This architecture combines the fine-grained modeling of pairwise interactions with scalable embedding computations, effectively capturing both local details and global structural context.

2.5. Continuous Tracking Paradigm and Cross-Correlogram

Recently, a method known as continuous tracking has been introduced to evaluate the temporal dynamics of the visual system. Originally developed in the context of eye-tracking studies to estimate reflexive and non-reflexive temporal responses (J. Mulligan et al., 2013), the method was later adapted for manual tracking tasks (Bonnen et al., 2015). To date, continuous tracking has been shown to provide a reliable approximation of the entire sensorimotor pipeline—from perception to action—while maintaining consistency with previous findings (Ambrosi et al., 2024; Barnett et al., 2025; Bonnen et al., 2017; Burge & Cormack, 2020). Moreover, it holds promise for probing more complex and higher-order visual-cognitive processes in a more efficient way.

During continuous tracking tasks, participants are typically instructed to use their eyes or a mouse to track the location (or value) of a target whose trajectory is governed by the temporal integration of independent white noise. Formally, the trajectory is defined as:

$$O(t) = \int_{i=0}^t W(t)dt, \quad (2-45)$$

$$\forall k, l, k \neq l, W(k) \perp W(l)$$

This temporal integration transforms uncorrelated noise $W(t)$ into a smoother trajectory $O(t)$, making it more tractable for human observers to follow. This property can be characterized by comparing the expected power spectral density (PSD) of $O(t)$ and $W(t)$. Although the Fourier transform of a stochastic process lacks a deterministic form, its statistical expectation can be analyzed using the Wiener–Khinchin theorem:

$$\begin{aligned} S_w(f) &= \mathfrak{F}\{\mathbb{E}[W(t)W(t+\tau)]\} \\ &= \mathfrak{F}\{\sigma^2\delta(\tau)\} = \sigma^2 \end{aligned} \quad (2-46)$$

Here, the $\mathfrak{F}\{.\}$ denotes the Fourier transform, $\mathbb{E}[.]$ is the expectation operator, and $\delta(\tau)$ is the Dirac delta function. This result shows that the PSD of white noise is flat, with energy evenly distributed across all frequencies.

In contrast, the power spectral density for integrated white noise $O(t)$ is given by:

$$S_o(f) = \left| \mathfrak{F}\left\{\int_{i=0}^t W(t)dt\right\} \right|^2 \quad (2-47)$$

$$= |\mathfrak{F}\{W * u(t)\}|^2 \quad (2-48)$$

Where $u(t)$ is a Heavyside step function:

$$\begin{aligned} u(t) &= \begin{cases} 1, & t \geq 0 \\ 0, & t < 0 \end{cases} \\ &= |\mathfrak{F}\{W(t)\} \cdot \mathfrak{F}\{u(t)\}|^2 \end{aligned} \quad (2-49)$$

Substituting from equation ((2-46), we obtain: (2-50)

$$S_o(f) = \sigma^2 \cdot \left| \frac{1}{i2\pi f} \right|^2 = \frac{\sigma^2}{4\pi^2 f^2}$$

This reveals that the power spectrum of the integrated white noise decays quadratically with frequency, emphasizing lower frequencies. Consequently, the resulting trajectory $O(t)$ is smoother and dominated by low-frequency components, which facilitates human tracking compared to raw white noise $W(t)$.

In studies utilizing continuous tracking paradigms, the cross-correlogram (CCG) is a classical tool for estimating the temporal dynamics of the visuomotor system. The CCG computes the normalized cross-correlation between the velocity profiles of the target trajectory $(\frac{dT(t)}{dt})$ and the participant's response $(\frac{dR(t)}{dt})$ averaged across multiple trials. The continuous-time formulation is given by:

$$CCG(\tau) = \mathbb{E}\left[\frac{Cov_{dR \cdot dT}(\tau)}{\sqrt{Cov_{dR \cdot dR}(0)Cov_{dT \cdot dT}(0)}}\right] \quad (2-51)$$

where the cross-covariance is defined as:

$$Cov_{A \cdot B}(\tau) = \int_{t=-\infty}^{\infty} A(t) \cdot B(t - \tau)dt$$

In discrete form, the cross-covariance becomes:

$$Cov_{A-B}(\tau) = \begin{cases} \sum_{i=1}^{n-\tau} \Delta A_{i+\tau,k} \Delta B_{i,k}, & \tau \geq 0 \\ \sum_{i=1}^{n+\tau} \Delta A_{i,k} \Delta B_{i-\tau,k}, & \tau < 0 \end{cases} \quad (2-52)$$

It can be shown that the CCG approximates the system's linear temporal dynamics, and if the system is linear and time-invariant (LTI), the CCG is equivalent to the system's impulse response function.

Let us consider an LTI system, L characterized by an impulse response $h(t)$. The response $R(t)$ to an input signal $I(t)$ is given by the convolution:

$$R(t) = L(I(t)) = I(t) * h(t) \quad (2-53)$$

Assuming the presence of additive system noise $\epsilon(t)$, the actual response becomes:

$$R(t) = I(t) * h(t) + \epsilon(t) \quad (2-54)$$

Assuming the input $I(t)$ is generated by integrating white noise $W(t)$, i.e., $I(t) = \int_0^t W(s)ds$.

We aim to prove that the CCG is equivalent to the system's impulse response.

Let $*$ denote convolution operation, $\otimes$ denote normalized cross correlation.

$$\begin{aligned} CCG(\tau) &= \mathbb{E}[dR(t) \otimes dI(t)] \\ &= \mathbb{E}[d(I(t) * h(t) + \epsilon(t)) \otimes dI(t)] \end{aligned} \quad (2-55)$$

$$= \mathbb{E}[(dI(t) * h(t) + d\epsilon(t)) \otimes dI(t)] \quad (2-56)$$

$$= \mathbb{E}[dI(t) \otimes dI(t) * h(t) + d\epsilon(t) \otimes dI(t)] \quad (2-57)$$

Given that $dI(t)$ is white noise, its autocorrelation is a delta function $\delta(\tau)$ (2-58)

$$= \mathbb{E}[\delta(\tau) * h(t)] + \mathbb{E}[d\epsilon(t) \otimes dI(t)]$$

$\because$ we assume noise independent of signal

$$\therefore \epsilon(t) \perp I(t) \rightarrow d\epsilon(t) \perp dI(t) \quad (2-59)$$

$$CCG(\tau) = h(t) + 0$$

Therefore, under the assumptions of linearity, time invariance, and additive uncorrelated noise, the cross-correlogram provides a direct estimate of the system's impulse response function $h(t)$

This property holds for any white noise process, regardless of its underlying probability distribution, as long as the noise has the following characteristics:

1. Zero mean and constant variance: The process $W(t)$ satisfies $\mathbb{E}[W(t)] = 0$ and $\mathbb{E}[W(t)W(t + \tau)] = \sigma^2 \delta(\tau)$.
2. Uncorrelated across time: Temporal independence implies a flat (i.e., constant) power spectral density.
3. Finite second-order moment: Ensuring the Fourier transform of the autocorrelation function exists in the L^2 sense.

Under these conditions, the Wiener–Khinchin theorem guarantees that the expected power spectral density remains constant for $W(t)$ and the convolution-based derivation of CCG as a linear impulse response holds regardless of the distribution type—Gaussian, uniform, Laplacian, etc. Therefore, the derivation of the cross-correlogram as a proxy for system dynamics is robust across a wide class of white noise inputs, making it a flexible and general tool for probing linear temporal characteristics in perceptual or motor systems.

Even if the input signal $I(t)$ exhibits temporal autocorrelation (i.e., is no longer generated by integration of white noise), the system’s linear response $h(t)$ can, in principle, still be recovered by deconvolution—that is, by inverting the convolution with the known autocorrelation structure of $dI(t)$. This follows from the relationship $CCG = h(t) * (dI(t) \otimes dI(t))$ when the temporal independence assumption is violated. However, such inversion presumes that the system remains linear and time-invariant across all frequency components. More critically, when the input contains structured temporal dependencies, it may engage nonlinear or adaptive mechanisms in the perceptual system, leading to context- or history-dependent responses that violate the assumptions of linearity. In such cases, the cross-correlogram may no longer reflect a stable impulse response, and the interpretation of system dynamics must therefore be treated with caution.

Chapter 3. Feature Invariant Process of Temporal Segmentation

To initiate the series of studies investigating cross-attribute temporal segmentation, this chapter first examines whether there exists a feature-invariant process specifically supporting temporal segmentation across different visual attributes. Building on the theoretical foundation established in the preceding chapters, we focus on whether temporal coherence—regardless of the specific visual attribute such as luminance, color, or motion—can serve as a general organizing cue for perceptual grouping.

3.1. Introduction

The human visual system uses the spatiotemporal structure of visual signals for visual segmentation and grouping. For example, the Gestalt principle of common fate states that objects moving at the same speed and direction tend to be grouped (Wertheimer, 1923). Furthermore, several studies suggest that visual attributes with similar temporal dynamics are grouped perceptually across spatial locations. Alais et al. (1998) demonstrated that synchronous alternation of the luminance contrast of four Gabor elements facilitated the integration of local Gabor motions into a global motion. Sekuler & Bennett (2001) made a checkerboard pattern composed of numerous cells, each sinusoidally modulated in luminance, and found that target cells with temporally phase-shifted luminance modulation were perceptually segregated. Accumulating research has shown that temporal synchrony (i.e., simultaneous alternation of visual attributes) causes a sense of grouping, whereas asynchrony causes a sense of segregation.

These phenomena are observed for a wide range of visual attributes, such as luminance, spatial frequency (Guttman et al., 2005), orientation (S.-H. Lee & Blake, 1999), and motion direction (Kandil & Fahle, 2004; Lee & Blake, 1999; Usher & Donnelly, 1998). Furthermore, Guttman et al. (2005) tested mixed-attribute conditions in which asynchrony was produced between two attributes selected from orientation, spatial frequency, spatial phase, and contrast. The results showed that participants could segregate an area from the background by temporal asynchrony, even when asynchrony was produced between different attributes. This suggests that spatial structure can emerge from a temporal structure abstracted from dynamic changes in any attribute

(Guttman et al., 2005). The attribute-invariant segmentation suggests that the computational mechanism of the asynchrony-based segregation may consist of the following stages (Figure 3-1): (1) separate attribute-specific channels that encode stimulus changes in a variety of attributes, (2) integration of attribute-specific signals into attribute-invariant signals, and (3) a universal process that compares attribute-invariant temporal signals to detect asynchrony across areas.

This study aims to test the generality of the findings by Guttman et al. (2005) and obtain insights into the underlying computational mechanism of asynchrony-based segregation. Since Guttman et al. (2005) only used changes in achromatic stationary attributes for mixed attribute tests, we used color changes and motion direction changes in addition to luminance changes. The visual temporal response is known to be slower for color changes (Burr & Morrone, 1993; Mullen et al., 2010) and motion changes (Werkhoven et al., 1992) compared to luminance changes. In addition, physically synchronous changes in color and motion direction can be seen as asynchronous (Moutoussis & Zeki, 1997). One factor suggested to produce this perceptual asynchrony is a difference in modulation waveform between color and motion direction (Nishida & Johnston, 2002). These differences may affect asynchrony-based spatial segregation.

To examine segmentation performance systematically based on temporal asynchrony involving luminance, color, motion direction, and their combinations, we used a square stimulus divided into four quadrants, each containing multiple patch elements. The elements changed one of the attributes at regular intervals, with the elements in one quadrant being phase delayed by 90° relative to the rest. By changing the temporal frequency of attribute changes and evaluating the ability of the visual system to discern the target from non-target quadrants at each temporal frequency, we derived the temporal frequency tuning of each stimulus condition. The similarity in the tuning characteristics would indicate the degree of affinity of the underlying process across different attributes, while the dissimilarity in the tuning characteristics would suggest where the differences may exist in the processing stream.

In what follows, we will report two experiments. Experiment 1-1 compared the asynchronous target detection performances among luminance, color, motion direction, and their combinations while using square wave as the waveform of stimulus modulation. The results showed that motion direction behaves differently from luminance and

color. Experiment 1-2 introduced triangular-wave modulation in addition to square-wave modulation, applying the two waveforms to luminance, color, spatial phase, and their combinations. The motivation for conducting Experiment 1-2 was to test a hypothesis that the differences among the attributes found in Experiment 1-1 might be due to the difference in the waveform. Since Experiment 1-1 is a subset of Experiment 1-2, we collected the data for the two experiments together to make the results of all the conditions directly comparable. We developed this data collection plan on the basis of the results of preliminary experiments.

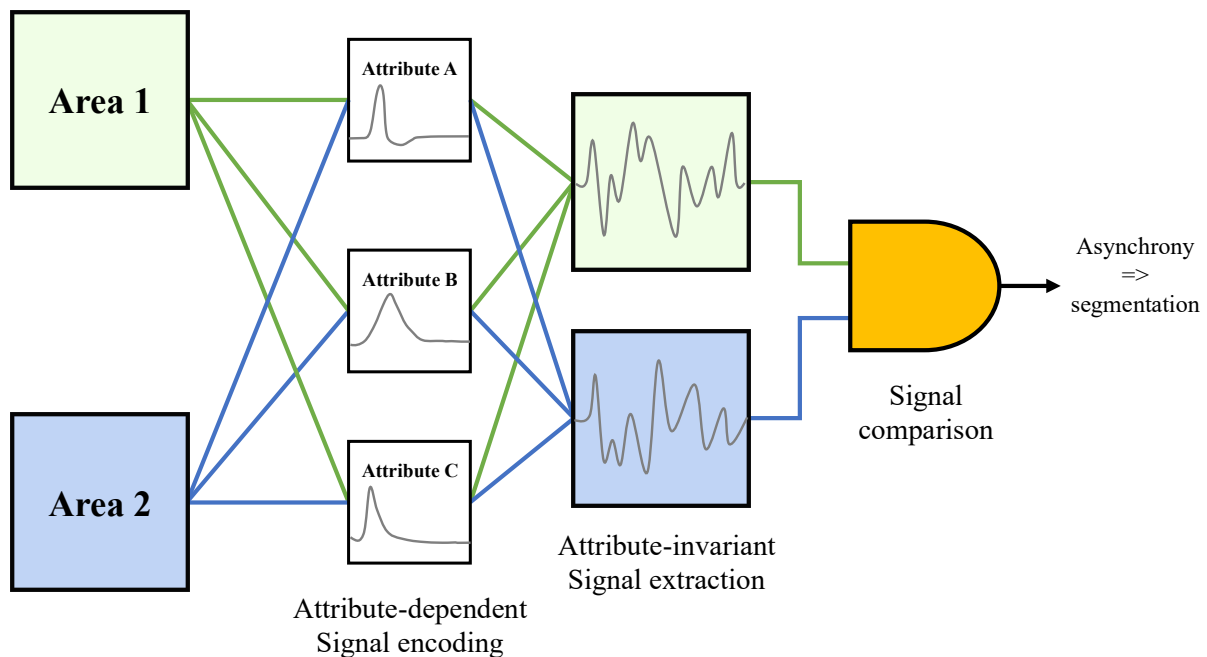


Figure3-1. A possible computational mechanism that can explain attribute-invariant segmentation based on temporal asynchrony.

First, the separated channels (with specific impulse response functions) encode the attribute changes at each location, then feed the outputs to the second stage, where attribute-invariant temporal signals are generated and compared across space. The spatial segmentation is produced when the comparison process detects asynchrony.

3.2. Experiment 1-1. Visual Segmentation on Luminance, Color, And Motion Direction

3.2.1. Method

3.2.1.1. Apparatus

The stimuli were displayed on a VIEWPixx /3D LCD monitor (VPixx Technologies, Saint-Bruno-de-Montarville, Canada) with a 1920×1080 pixels resolution and a 120 Hz refresh rate. The lowest, mean, and highest display luminance was 1.8, 48.4, and 96.7 cd/m^2 , respectively. The linear output of the luminance was calibrated using an i1Pro chromometer (VPixx Technologies) for each channel. Each pixel subtended 1.3 arcmin at a viewing distance of 70 cm. The participant was seated in a dark room with a chinrest stabilizing the head. The experiments were programmed using the Psychophysics Toolbox (Brainard, 1997) in MATLAB release 2022a (Math Works, Natick, MA, USA). The study protocol followed the ethical standards of the Declaration of Helsinki except for preregistration and was approved by The Ethics Committee of Kyoto University (KUIS-EAR-2020-003).

3.2.1.2. Participants

The experiment included two of the authors and two naïve participants (four males, mean age 24.67 years). The participants declared that they had normal or corrected-to-normal vision. All the participants were informed of the nature of the research and provided informed consent before the experiment started. Participants received monetary compensation after the experiment finished. A revised version of the minimum motion method (Anstis & Cavanagh, 1983; Cavanagh et al., 1987) was used to measure the equiluminant ratio (in terms of the system intensity value) between the red (R) and green (G) channels for each participant before the main experiment. During the procedure, participants saw a Gabor drifting rightward. The sinusoidal wave of the R and G channels is 180 degrees out-of-phase. Participants were required to adjust the ratio between the channels to minimize the sense of motion. The mean R: G ratio value was 0.58.

3.2.1.3. Stimuli

In each trial, a square stimulus was displayed for 1000 ms. The stimulus was divided into four quadrants, each composed of 8×8 tightly tiled $0.5^\circ \times 0.5^\circ$ patches. The patches alternated in luminance, color, or motion direction. Patches in the same quadrant made changes in the same attribute and with the same timing. The changes in one of the quadrants were delayed relative to the remaining three quadrants. We postulated that this target quadrant would be segregated perceptually.

When the luminance was modulated, the luminance profile of the patch was a Gaussian blob as follows (except under the cross-attribute conditions paired with motion direction; see below):

$$L(x, y, t; \sigma) = L_{\text{mean}} \cdot \left(1 + c(t; \epsilon, \xi, A_L, C_0) \cdot e^{-\frac{x^2+y^2}{2\sigma^2}} \right) \quad (3-1)$$

$$c(t; \epsilon, \xi, A_L, C_0) = \pm A_L \cdot \text{Sqr}(t; \epsilon, \xi) + C_0 \quad (3-2)$$

$$\text{Sqr}(t; \epsilon, \xi) = \text{SGN}(\sin(2\pi\xi t - \epsilon)) \quad (3-3)$$

$$\text{SGN}(\sin(2\pi\xi t - \epsilon)) = \begin{cases} 1, & \sin(2\pi\xi t - \epsilon) \geq 0 \\ -1, & \sin(2\pi\xi t - \epsilon) < 0 \end{cases} \quad (3-4)$$

Where x, y corresponded to the spatial location; t represents the time in seconds; σ modulates the width of the Gaussian profile, which was set as 6 arcmin; L_{mean} is the mean luminance; $c(t)$ is the temporal contrast modulation function, where ξ is the temporal frequency (TF) in hertz; and ϵ is the temporal phase shift, set as $\frac{1}{4\xi}$ for the delayed quadrant, or equivalent to a 90° phase delay; A_L determines the maximum modulation amplitude, which is set as 0.3; C_0 is a mean contrast, a constant drawn from the uniform distribution $[-0.7, 0.7]$. This term randomly shifts the waveform upward or downward, completely tiling the entire value range to exclude additional luminance cues. $\text{Sqr}(t)$ is the square wave function, which changed the sinusoidal function by adopting the sign function (annotated as $\text{SGN}(f(t))$). The sign function mapped all the positive real numbers and zero to positive one while mapping all the negative real numbers to negative one.

When the color was modulated, the luminance profile of the patch was also a Gaussian blob as follows (except under the cross-attribute conditions paired with motion direction; see below):

$$\begin{cases} R(x, y, t; \sigma) = L_{\text{mean}} \cdot \left(1 + c(t; \epsilon, \xi, A_c, R_0) \cdot e^{-\frac{x^2+y^2}{2\sigma^2}} \right) \\ G(x, y, t; \sigma) = -L_{\text{mean}} \cdot \left(1 + L_{\text{iso}} \cdot c(t; \epsilon, \xi, A_c, G_0) \cdot e^{-\frac{x^2+y^2}{2\sigma^2}} \right) \\ B(x, y, t) = 0 \end{cases} \quad (3-5)$$

where R, G, and B represent the three color channels and L_{iso} is the isoluminant index between R and G; A_c determines the maximum modulation amplitude and its sign for each temporal profile, either 0.5 or -0.5, was randomly assigned to each patch, R_0 and G_0 is the mean chromatic contrast, a constant drawn from the uniform distribution $[-0.5, 0.5]$, and $R_0 = G_0$. The constant is shared between R and G channel. $c(t; \epsilon, \xi, A_c, R_0)$ and $c(t; \epsilon, \xi, A_c, G_0)$ is the same function as equation (3-2). The only difference is changing A_L into A_c while C_0 into R_0 and G_0 respectively.

When the motion direction was modulated, the luminance profile of the patch was a Gabor function:

$$\begin{aligned} L(x, y, t; \sigma, \omega, \theta, \phi) \\ = L_{\text{mean}} \left(1 + \cos(2\pi\omega r - \phi(t; \epsilon, \xi, A_M, \phi_0)) \cdot e^{-\frac{x^2+y^2}{2\sigma^2}} \right) \end{aligned} \quad (3-6)$$

$$r = y\cos(\theta) + x\sin(\theta) \quad (3-7)$$

$$\phi(t; \epsilon, \xi, A_M, \phi_0) = \pm A_M \cdot \left[2 \left\lfloor 2 \left(t\xi - \left\lfloor t\xi + \frac{1}{2} \right\rfloor \right) \right\rfloor - 1 \right] + \phi_0 \quad (3-8)$$

where ω is the spatial frequency, set as one cycle per degree; θ is the spatial orientation, randomly drawn from the uniform distribution within the interval $[-\pi, \pi]$; $\phi(t)$ is the temporal modulation function for the spatial phase; $\left\lfloor t\xi + \frac{1}{2} \right\rfloor$ is the floor function outputs the closest natural number that is not larger than the input. A_M is the magnitude of shift per second, which is 90° or -90° , and ϕ_0 is the mean phase, a constant drawn from the uniform distribution, which ranges $[-90, 90]$. In the actual code, we implemented the last function using an equivalent integral form as follows:

$$\phi(t; \epsilon, \xi, A_M, \phi_0) = \pm A_M \cdot \left[\frac{4\xi}{\mathbb{F}} \sum_0^t \text{Sqr}(t; \epsilon, \xi) - 1 \right] + \phi_0 \quad (3-9)$$

$\mathbb{F}$ is the sampling rate of the monitor, which is 120 Hz. With this integral form, we can eliminate the artifact caused by the down-sampling from a continuous curve.

When the luminance or color was paired with motion direction in cross-attribute conditions, we used the Gabor function:

$$\begin{aligned} L(x, y, t; \sigma, \omega, \theta, \phi) \\ = L_{\text{mean}} \\ \cdot \left(1 + c(t; \epsilon, \xi, A_L, C_0) \cdot \cos(2\pi\omega t - \phi) \cdot e^{-\frac{x^2+y^2}{2\sigma^2}} \right) \end{aligned} \quad (3-10)$$

and the RGB profile in the color condition:

$$\begin{cases} R(x, y, t; \sigma) = L_{\text{mean}} \cdot \left(1 + c(t; \epsilon, \xi, A_C, R_0) \cdot \cos(2\pi\omega t - \phi) \cdot e^{-\frac{x^2+y^2}{2\sigma^2}} \right) \\ G(x, y, t; \sigma) = -L_{\text{mean}} \cdot \left(1 + L_{\text{iso}} \cdot c(t; \epsilon, \xi, A_C, G_0) \cdot \cos(2\pi\omega t - \phi) \cdot e^{-\frac{x^2+y^2}{2\sigma^2}} \right) \\ B(x, y, t) = 0 \end{cases} \quad (3-11)$$

where ϕ was fixed at a value randomly drawn from $[-\pi, \pi]$. $c(t; \epsilon, \xi, A_L, C_0)$ is the same formula as equation 3-1. $c(t; \epsilon, \xi, A_C, R_0)$, $c(t; \epsilon, \xi, A_C, G_0)$ are the same as equation (3-5

All the stimulus profiles were remapped to an 8-bit scale before being presented. Figure 3-2 shows a sample frame and waveform in the lower panel.

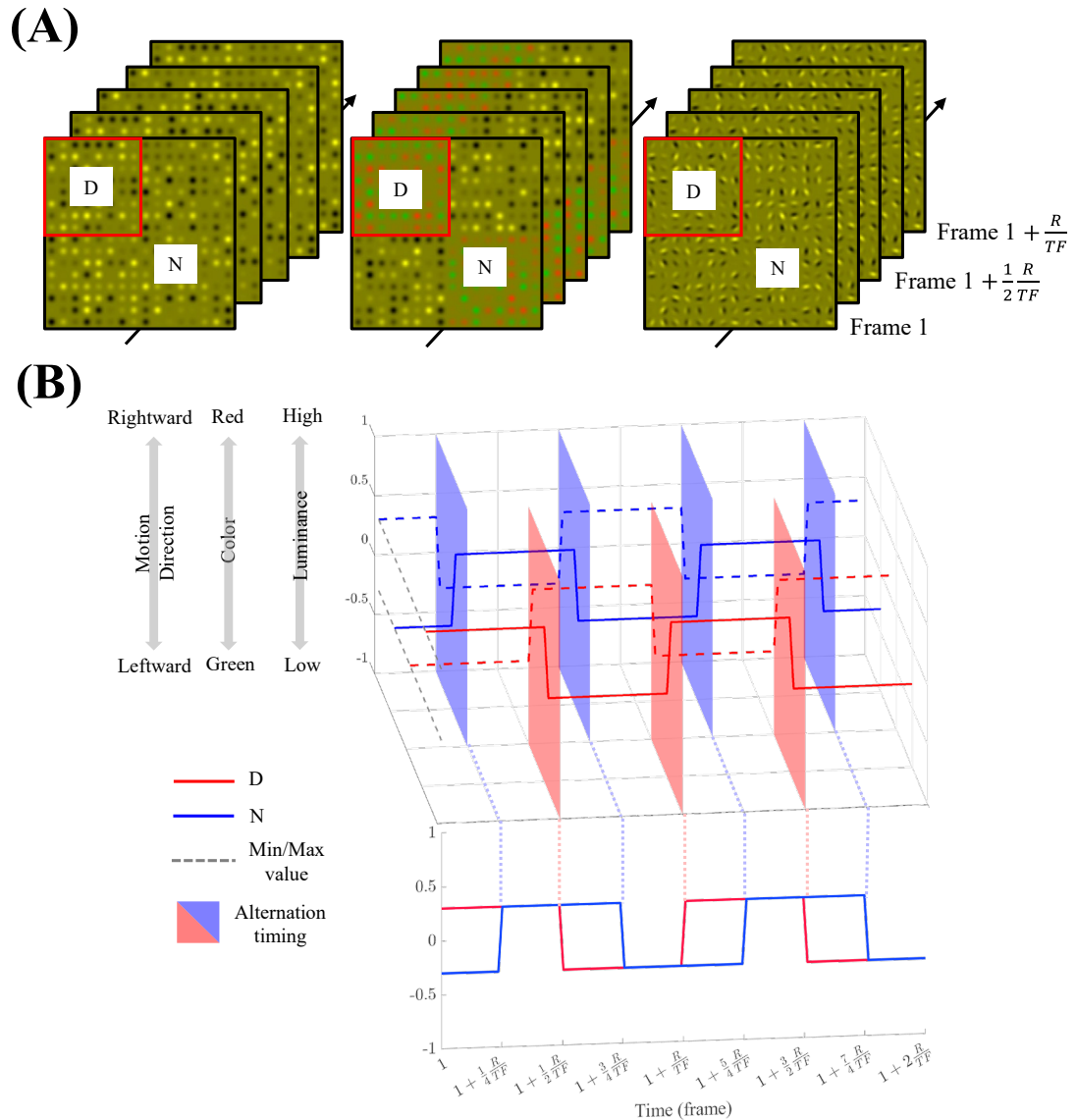


Figure 3-2. Stimuli for Experiment 1-1.

(A) Stimulus snapshots for within-attribute luminance (left), cross-attribute luminance and color (center), and cross-attribute luminance and motion direction (right). (B) The temporal modulation waveforms for areas D for delayed area (red) and N non-delayed area (blue). The modulation was a square wave and was temporally delayed for area D relative to area N. The polarity of modulation (solid or dashed line) was chosen randomly for each element in both areas. The red and blue planes indicate the frames in which the pattern was updated. R is the refresh rate of the monitor. The lower side, attached to the x-axis, shows the 2-D projection.

3.2.1.4. Design and Procedure

A method of constant stimuli with a four-alternative-forced-choice task was used. In each trial, participants viewed a 1-s stimulus clip and reported which quadrant was most visually segregated from the others by pressing the "Q," "W," "A," or "S" keys on the keyboard for left upper, right upper; left lower, or right lower quadrant, respectively. The location of the delayed quadrant was randomized across the trials. No

fixation point was given. Three attributes were used: luminance (Lum), color (Col), and motion direction (Dir). There were two different attribute presentation conditions: within-attribute and cross-attribute. Only one of the attributes was modulated in the within-attribute condition, whereas two of the three attributes were modulated in the cross-attribute condition. In the latter case, one attribute changed in two diagonal quadrants, while another attribute changed in the remaining two quadrants. The attribute changing in the target quadrant was chosen randomly. There were six stimulus conditions in total: three within-attribute conditions and three cross-attribute conditions (Lum-Col, Lum-Dir, and Col-Dir).

The response was measured at 11 temporal frequencies ranging from 0 to 10 Hz in 1 Hz steps. The stimulus condition was fixed in a session, while the temporal frequency conditions were mixed. Each participant was required to complete all the conditions. The session order was randomized for each participant. There were 30 repetitions for each stimulus condition, temporal frequency, and participant combination. Six stimulus conditions for Experiment 1-1 were collected together with 15 stimulus conditions for Experiment 1-2. In total, $21 \text{ (stimulus condition)} \times 11 \text{ (temporal frequency)} \times 30 \text{ (trial)} = 6930$ trials were run by each participant. It took about three to four hours to finish Experiment 1-1 and 1-2.

3.2.2. Result

We computed the proportion of the correct responses (i.e., the responses choosing the delayed quadrant) to evaluate segregation performance. The grand average of all participants' data was used to inspect the general performance. A bootstrap procedure is applied to each temporal frequency for estimating the confidence interval. Within each repetition, we resampled the four participants' proportions correct with replacement, then calculated one mean value on the sampled proportion correct. The process was repeated 100,000 times, and the confident area was defined as the range between the 2.5% quantile and the 97.5% quantile of the repetitions.

For conditions including only color and luminance (within-attribute luminance, within-attribute color, and cross-attribute luminance and color; the first column of Figure 3-3, the grand average showed high performance for all the conditions unless the temporal frequency was very high. The segmentation accuracy was perfect at times for the two within-attribute conditions. The cross-attribute condition also reached 100%

accuracy for some observers. There was a minor difference in the temporal tuning among the three conditions. The frequency range at which the performance was significantly higher than the guess rate (25%) was broader for luminance (1–7 Hz) than for color (1–5 Hz) or for their combination (1–5 Hz).

By contrast, the segmentation accuracy was low for conditions including motion direction (the second and third columns of Figure 3-3. The maximum target detection accuracy was around 60% for the within-attribute motion direction, unlike luminance and color. Additionally, the performance obtained under cross-attribute conditions rarely surpassed the guess rate significantly, indicating an inability to accomplish the task.

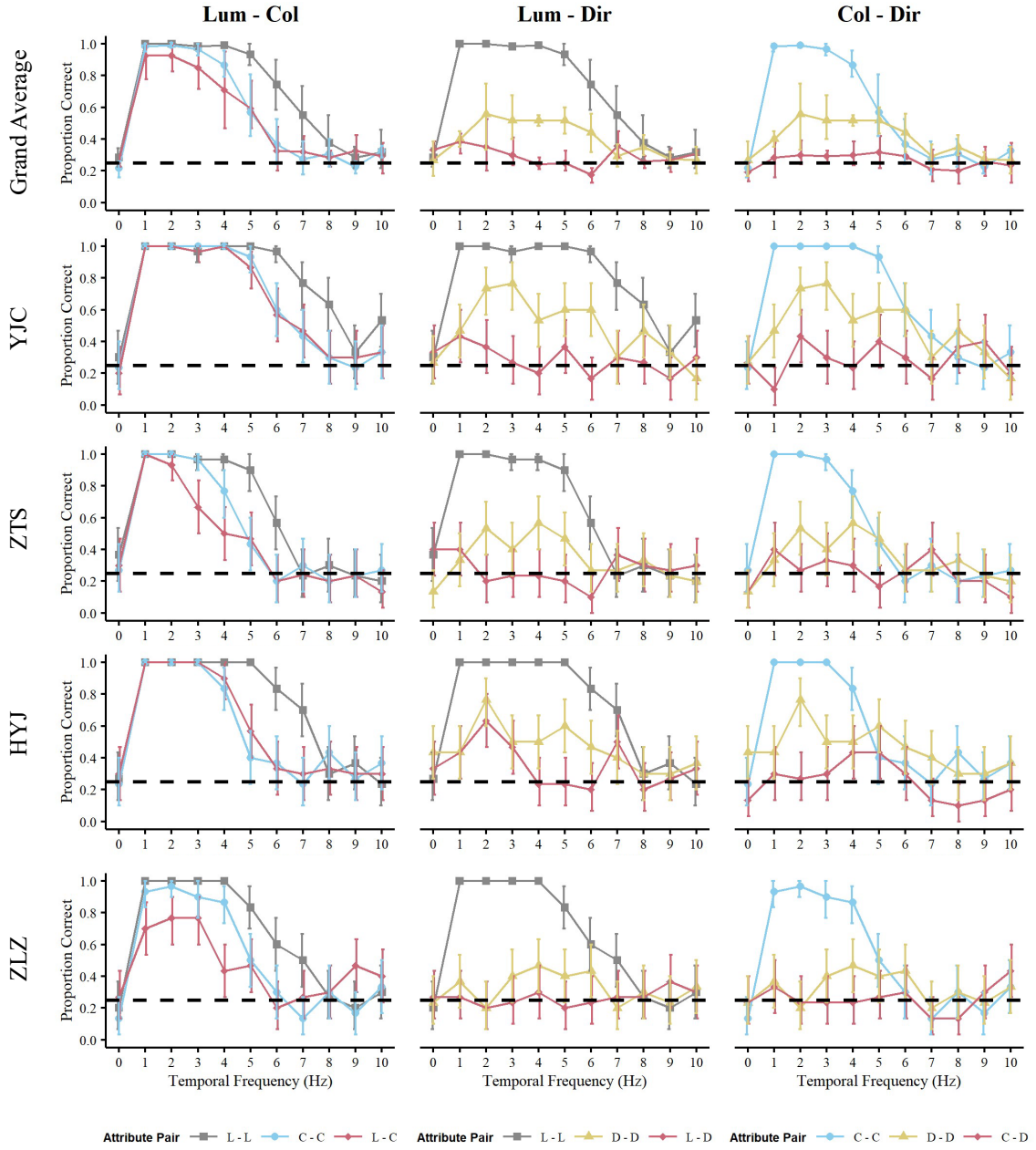


Figure 3-3. Proportion correct as a function of the temporal frequency for each attribute pair.

The first row shows the grand average curves over 4 participants; the second to fifth rows show individual curves separately. Different attribute pairs are shown with distinct colors and symbols. The red line with diamonds corresponds to the cross-attribute condition, while the other color lines are within-attribute conditions: the gray line with squares represents luminance, the blue line with circles represents color, and the yellow line with triangles represents motion direction. Luminance is abbreviated as Lum/L, Color as Col/C, and motion direction as Dir/D. The error bar for each point represents the centralized 95% confidence interval estimated using the bootstrapping procedure in grand average, while the error bar for each participant is the centralized 95% interval of binomial distribution. The black dashed line indicates the guess rate (0.25).

3.2.3. Interim discussion

Experiment 1-1 measured the frequency tuning characteristics for visual segmentation across luminance, color, and motion direction to investigate whether the performance was invariant regardless of the attribute. In the within-attribute condition, the result indicated that participants could segregate the delayed quadrant regardless of the attribute, consistent with previous studies (Alais et al., 1998; Sekuler & Bennett, 2001; Usher & Donnelly, 1998). However, the performance was significantly worse for motion direction than for luminance or color. Furthermore, in the cross-attribute condition, the segmentation task was possible with the luminance-color pair but nearly impossible for the two pairs that included motion direction. The heterogeneity in segmentation performance among different stimulus conditions indicates a limitation of the simple attribute invariance hypothesis and a unique characteristic of motion direction.

One possible factor explaining poor performances in conditions related to motion direction is the slow temporal response of motion processing. For example, the sensitivity to detect motion acceleration suggests that the temporal integration of the velocity signal is at least 100 ms (Werkhoven et al., 1992). However, this factor does not explain why we found very poor performance for cross-attribute conditions including motion direction, even at low temporal frequencies. In addition, Maruya et al. (2013) found that the upper limit to detect the synchrony of a pair of motion direction alternations could be high (> 10 Hz) enough to cover the frequency range we tested. Given these findings, the observed poor performances in conditions related with motion direction should not be ascribed solely to the low temporal resolution of motion processing.

One may wonder why the performance of the within-attribute motion condition was not as good as Maruya et al. (2013) reported. This study showed high temporal limits for within-motion synchrony judgments only when the compared motion pair had a common component vector. The temporal limit dropped to ~ 3 Hz when no common component existed. This drop is likely to reflect the temporal characteristics of the motion change comparison stage rather than that of the initial change detection stage. In our stimulus configuration, the elements' motion directions were random, so they did not have a common vector component. Lacking common components may be one reason the performance of the within-attribute motion condition was not good in the

present study. By somehow increasing the spatial coherence of local motion directions, we might be able to improve the performance of asynchrony-based segregation and find a different temporal characteristic. This is an issue left for further study. In what follows, we will focus on the results found with our specific motion stimuli in which elements had random orientations and directions and made 180-deg motion reversals.

Cross-attribute conditions including motion direction (luminance vs. motion, or color vs. motion) showed very poor segregation performance. Regarding this finding, we hypothesized that the mechanism underlying a color motion asynchrony (Moutoussis & Zeki, 1997) may also play some role in synchrony-based grouping and asynchrony-based segregation. In this cross-attribute timing illusion, physically synchronous alternations in color and motion direction are perceived as asynchronous, with motion direction changes being apparently delayed relative to color changes. This apparent delay could impair synchrony-based grouping between motion changes and color (or luminance) changes. One explanation for the color-motion asynchrony effect is waveform mismatching (Nishida & Johnston, 2002; for a review of their time marker theory, Fujisaki et al., 2012). The modulation of color signals showed instantaneous changes, maintaining a value for a brief period and then swiftly shifting to another value. By contrast, the modulation of motion direction gradually shifted the spatial phase in one direction for a period, followed by a gradual shift in the opposite direction. Forced matching between a first-order change (a change of a non-dynamic attribute that can be defined at a single measurement in time, such as a change in luminance or color) and a second-order change (a change in the direction of change, such as a reversal of motion direction or luminance change direction) would result in mismatching in time. Following this hypothesis, an effect similar to color-motion asynchrony was observed for a variety of attribute pairs between first-order change (square-wave attribute change) and second-order change (triangular-wave attribute change) (Nishida & Johnston, 2002). We suspected that the dissimilar tuning curve of the motion direction and the other attributes in Experiment 1-1 may also be due to the different alternation patterns of the signals. Therefore, Experiment 1-2 investigated the impact of the waveform type.

3.3. Experiment 1-2. Visual Segmentation on Different Waveforms

To compare the waveform effect among the attributes systematically, we applied square wave modulation (making constant interval first-order changes) and triangular

wave modulation (making constant interval second-order changes) to the three attributes: luminance, color, and spatial phase. Note that application of a triangular wave modulation to the spatial phase shift was equivalent to Experiment 1-1's motion direction change condition. If the waveform induced different segmentation abilities, we would expect three results from Experiment 1-2. First, when using square wave modulation of the spatial phase, the tuning curve should be similar to that for luminance and color. Second, when using triangular wave modulation for luminance and color, the tuning curve should be similar to that for motion direction. Third, when mixing square wave modulation and triangular wave modulation, the tuning curve should be similar to the luminance-motion direction or the color-motion direction pair. We hypothesized that if asynchrony-based segmentation is attribute-invariant, we would find similar tuning curves regardless of the attribute and the combination after equalizing the waveform among luminance, color, and spatial phase.

3.3.1. Method

The experiment included two authors and two naïve participants (four males, mean age 24.67 years), identical to Experiment 1-1. The method was identical to that of Experiment 1-1, except that two temporal modulation functions were introduced: square wave ($Sqr(t)$) and triangular wave ($Tri(t)$). The two functions with aligned changing points were defined as follows (a demonstration of the waveform is shown in Figure 3-4):

$$Sqr(t; \epsilon, \xi) = SGN(\sin(2\pi\xi t - \epsilon)) \quad (3-12)$$

$$Tri(t; \epsilon, \xi) = \frac{4\xi}{\mathbb{F}} \sum_0^t Sqr(t; \epsilon, \xi) - 1 \quad (3-13)$$

There were six different attribute conditions: luminance (Lum), color (Col), spatial phase (Pha), and their three combinations. For each, three waveform configurations were introduced: all square wave, all triangular wave, and mixed waves. In the mixed wave condition, the quadrants on the same diagonal followed identical temporal profiles, whereas those on the other diagonal followed the other temporal profile. For example, the upper-left and lower-right quadrants followed a square wave, while the lower-left and upper-right quadrants followed a triangular wave. The square and

triangular waves were considered synchronous when the timing of the change in the value of the square wave and that of the slope of the triangular wave matched in time. In total, there were 21 stimulus conditions: nine within-attribute conditions (three attribute conditions $\times$ three waveform conditions) and twelve cross-attribute conditions (three attribute conditions $\times$ four waveform conditions). Cross-attribute conditions had one more mixed wave condition than within-attribute conditions because each attribute could be a square or triangular wave. For example, in the Lum-Col conditions, a square wave was given to Lum, while a triangular wave to Col, or vice versa.

Since six stimulus conditions were common to Experiment 1-1, we collected the data for the whole experiment together. The results for the six common conditions were based on the same data.

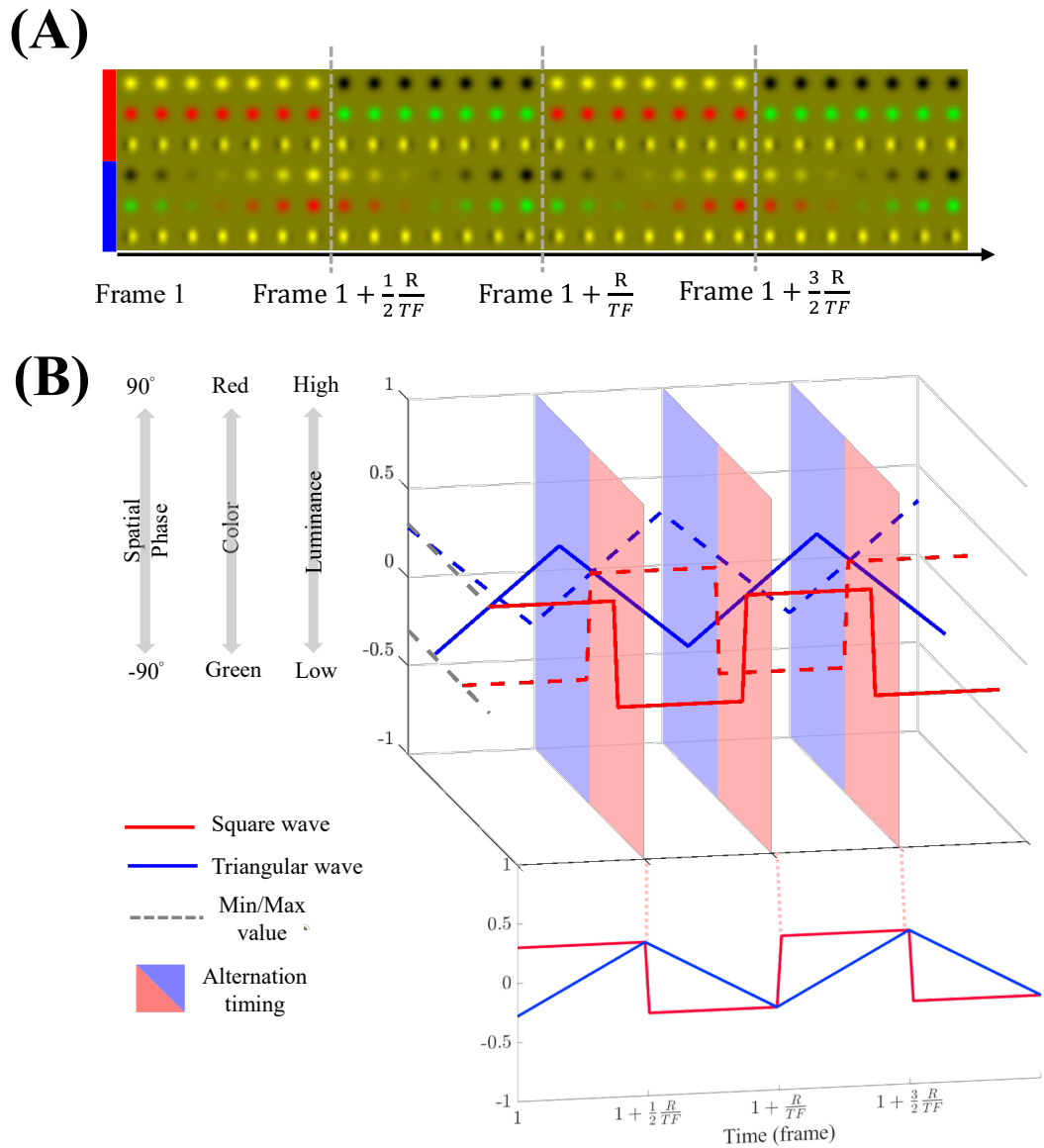


Figure 3-4. Experiment 1-2 used square and triangular waves.

(A) shows six examples of time series for one element in the stimuli (30 frames for each). The top three rows follow the square-wave modulation, and the bottom three follow the triangular-wave modulation. The corresponding attributes are luminance, color, and spatial phase from top to bottom. (B) shows the temporal modulation waveforms for square (red) and triangular (blue) waves. The square and triangular waves are aligned in terms of the simultaneity for the update timing of the square wave and the slope alternate timing of the triangular wave. The red and blue planes indicate the timing. The polarity of modulation (solid line or dashed line) is chosen randomly for each element in the stimuli. R is the refresh rate of the monitor. The lower side, attached to the x-axis, shows the 2-D projection.

3.3.2. Result

We will use L, C, and P to represent luminance, color, and spatial phase, respectively; 1 represents a square wave (first-order change), and 2 represents a triangular wave (second-order change). The top two rows of Figure 3-5 compare the performances

for two within-attribute conditions and the corresponding cross-attribute condition separately for each attribute combination. All the individual graph please refer to the 9.1

All square-wave conditions shown in the first row were identical to those in Figure 3-3, except spatial phase changes replaced motion direction changes. All within-attribute conditions resulted in a nearly perfect performance (proportion correct close to 100%) when the frequency was 1-5Hz for luminance, 1-3 Hz for Color, and 1-5 Hz for spatial phase. All cross-attribute conditions (L1-C1, L1-P1, and C1-P1) also showed good performances at lower temporal frequencies, but the effective frequency range was shrunk: 1-5 Hz for L1-C1 and C1-P1, and 1-4Hz for C1-P1. In general, the pattern of results was more invariant against attribute variation compared with the results of Experiment 1-1, especially for cross-attribute conditions, including the spatial phase. However, they were not perfectly attribute-invariant. For within-attribute conditions, the upper temporal limit was higher for spatial phase and luminance than for color. For cross-attribute conditions, the performance was generally worse than for the within-attribute conditions. Among the three conditions, the performance was slightly worse for L1-P1 and C1-P1 than for L1-C1.

When all the waveforms were switched to triangular waves (the second row of Figure 3-5), the segmentation task became much more difficult. The performances for the within-attribute conditions (L2-L2 and C2-C2) were far from perfect, even at the best temporal frequency, comparable to the performance of the P2-P2 condition (identical to the motion direction change condition in Experiment 1-1). The lowering of the performance can also be shown by the shrinkage of the effective frequency range, 2-6Hz in luminance, 1-3 Hz in color, and 1-6 Hz in spatial phase. The performances for the cross-attribute conditions (L2-C2, L2-P2, and C2-P2) were even poorer. The effective frequency range was reduced to 2-3 Hz for L2-C2, 3-5 Hz for L2-P2, and 1-4 Hz for C2-P2. In general, the tuning curves were not very different among different attribute conditions, as Experiment 1-1 has shown, although they were not perfectly attribute invariant, as we found with the square-wave conditions.

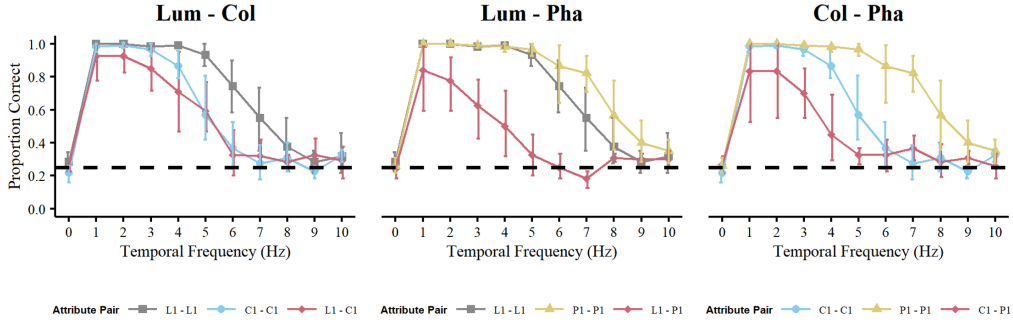
The effects of the waveform can be seen more clearly in the bottom two rows of Figure 3-5. In the within-attribute condition (the third row), participants consistently showed the best segmentation performance with square waves and worse performance with triangular waves. The accuracy fell below the guess rate when the square and triangular waves were mixed. A similar result was obtained for the cross-attribute conditions (the fourth row), except that the overall performance was worse than the

within-attribute conditions, and the tendency of the accuracy to fall below the guess rate for the mixed wave conditions was unclear.

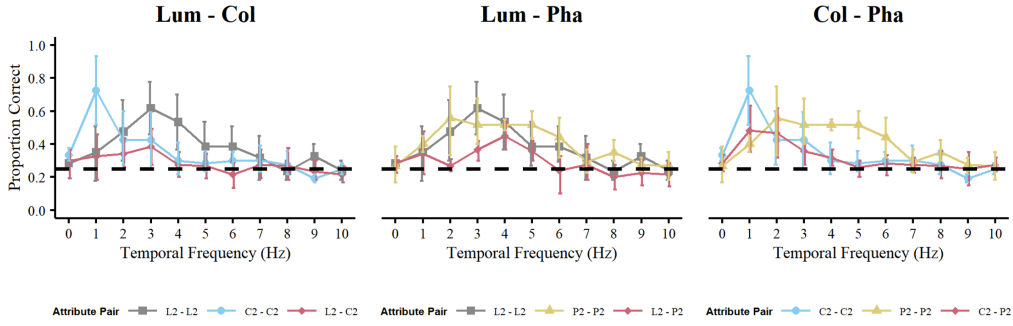
Although the participants could not detect the physically delayed target quadrant in the mixed wave conditions, they often made the inverse response, i.e., the responses choosing the quadrant located diagonal to the target (Figure 3-6); the detailed values of mixed-wave condition were shown in Table 3-1. Although we defined the correct choice as the quadrant with a physical asynchrony in stimulus change relative to the other three quadrants regardless of other stimulus manipulations, the asynchronous quadrant may not be the one that appeared most segregated from the others for cross attribute/order conditions. In cross attribute/order conditions, let us assume two quadrants are A, and the other are B, i.e., {A, A, B, B}. If temporal synchrony is given to one of A's, then the quadrant set can be described as {A⁺, A⁻, B⁻, B⁻}. Since B⁻ is the majority, the question is whether the observer chooses A⁺ or A⁻. The response for A⁺ is the correct response, while that for A⁻ is the inverse response.

As shown in Fig. 6, the proportion of the inverse responses was below the guess rate for the conditions where participants could make correct choices toward the target. However, when the waveform was not matched (i.e., mixed wave), the proportion of choosing the inverse response rose above the guess rate. The same observations as the cross-attribute conditions in the correct response, the proportion of choosing the inverse response also decreased. Besides, the frequency range for the inverse response was similar to that of the correct responses. For within-attribute conditions, the range was 1-8 Hz for luminance, 1-3 Hz for Color, and 1-8 Hz for spatial phase. For cross-attribute conditions, the range was not very clear.

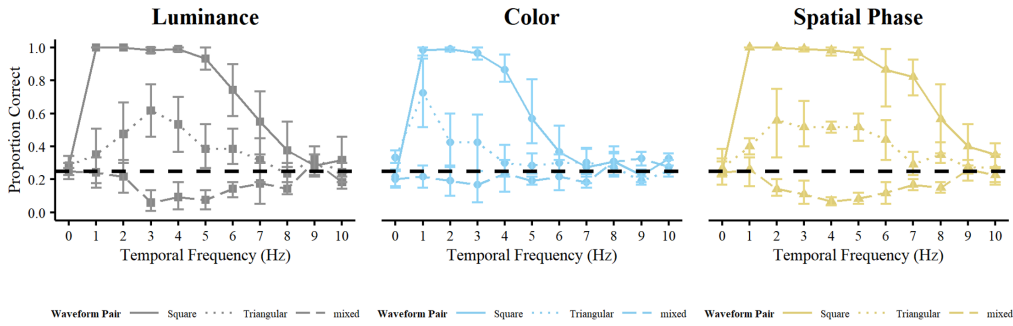
Within- vs. cross-attributes, square waves



Within- vs. cross-attributes, triangular waves



Within-attribute, all waveform combinations



Cross-attribute, all waveform combinations

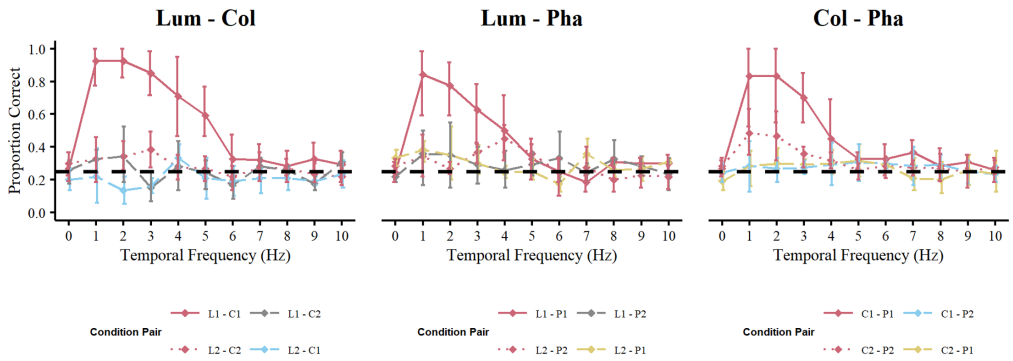
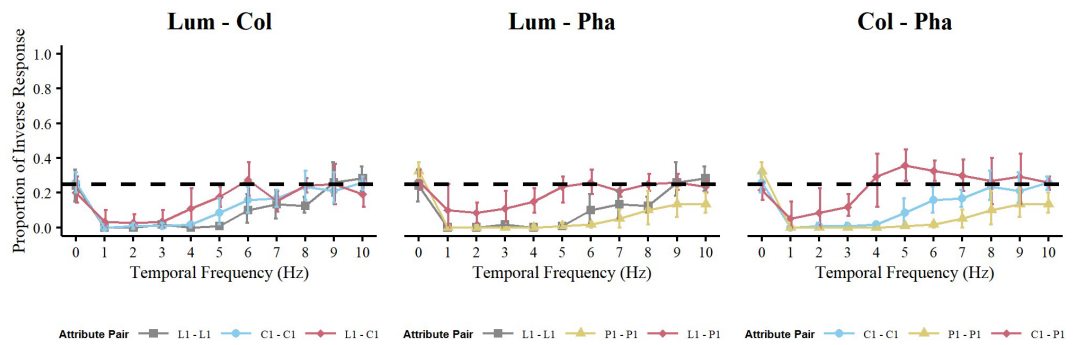


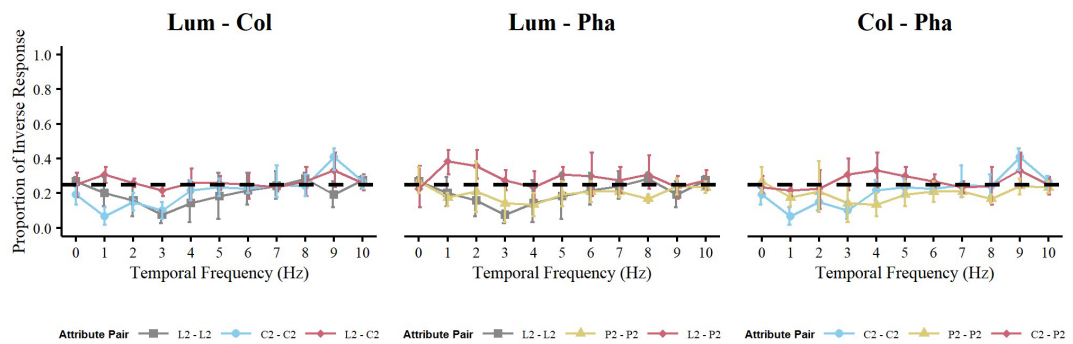
Figure 3-5. Proportion correct as a function of the temporal frequency for each attribute and waveform pair.

Rows 1 and 2 compare the performance among within- and cross-attribute conditions under the all-square-wave and all-triangular-wave results, respectively, with distinct colors and symbols representing attribute pairs. Luminance is abbreviated as Lum/L, Color as Col/C, and phase as Pha/P; 1 stands for square wave modulation, and 2 stands for triangular wave modulation. Rows 3 and 4 compare performance among waveform pairs under within-attribute and cross-attribute conditions, respectively. Error bars represent the 95% confidence interval, and the black dashed line indicates the guess rate (0.25).

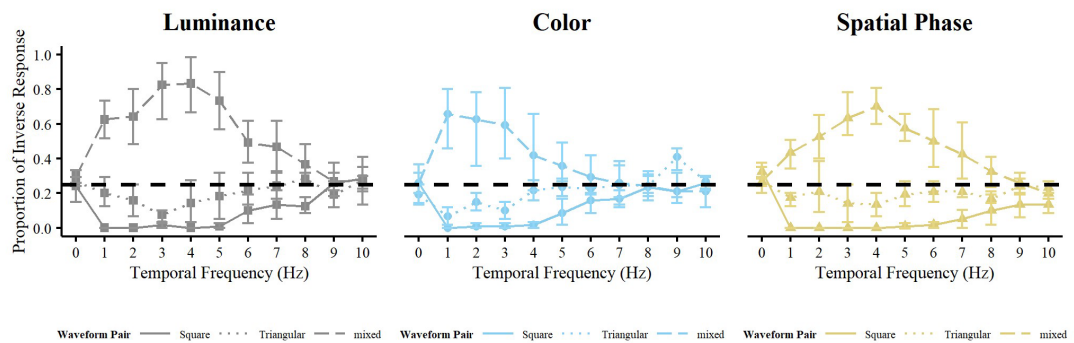
Within- vs. cross-attributes, square waves



Within- vs. cross-attributes, triangular waves



Within-attribute, all waveform combinations



Cross-attribute, all waveform combinations

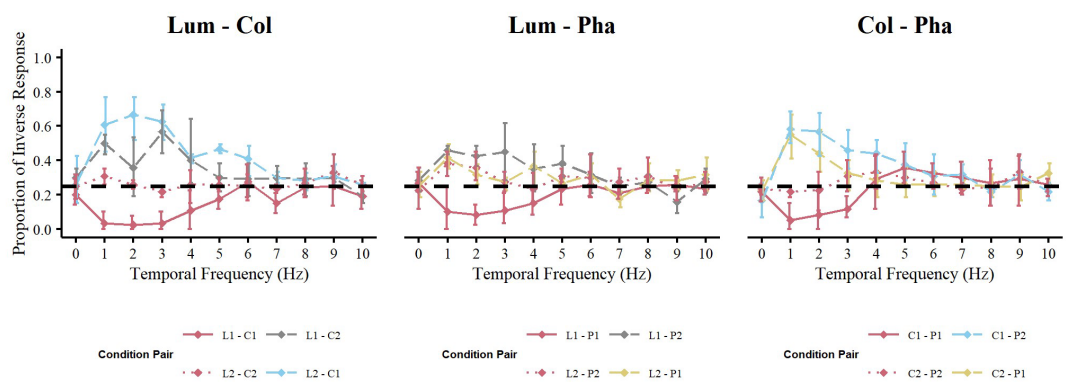


Figure 3-6. The proportion of inverse responses as a function of the temporal frequency of each attribute and waveform pair.

Table 3-1. Proportion of choosing target or inverse response in mixed-wave conditions, within-attribute

Proportion of selection		0 Hz	1 Hz	2 Hz	3 Hz	4 Hz	5 Hz	6 Hz	7 Hz	8 Hz	9 Hz	10 Hz
Luminance	Target	0.25	0.24	0.22	0.06	0.09	0.08	0.14	0.18	0.14	0.31	0.18
	Inverse	0.30	0.63	0.64	0.83	0.83	0.73	0.49	0.47	0.37	0.27	0.28
color	Target	0.20	0.22	0.19	0.17	0.23	0.19	0.22	0.18	0.31	0.33	0.28
	Inverse	0.26	0.66	0.63	0.59	0.42	0.36	0.29	0.26	0.24	0.21	0.21
Spatial	Target	0.24	0.26	0.14	0.11	0.07	0.08	0.12	0.17	0.15	0.26	0.23
Phase	Inverse	0.27	0.43	0.53	0.63	0.70	0.58	0.50	0.43	0.33	0.26	0.20

Note. A bold style annotates the value significantly higher than the guess rate.

3.4. General Discussion

3.4.1. Effects of waveform type or the order of changes

Experiment 1-1 showed that the target detection performance was worse for motion direction changes than for luminance and color changes. Experiment 1-2 showed that the target detection performance was generally worse for second-order changes produced by triangular-wave modulations than for first-order changes produced by square-wave modulations. Regardless of the attribute and the within- or cross-attribute condition, the performance was best when there was asynchrony between first-order changes, worse when there was asynchrony between second-order changes, and worst (close to the chance level) when there was asynchrony between first- and second-order changes.

The poor performance when mixing the first- and second-order changes can be explained by the waveform (dis)similarity. If the visual system detects (a)synchrony based on the cross-correlation-like computation of sensory signals that preserve the effect of this phase shift, a timing-aligned square and triangular waves will be perceived as asynchronous. As shown in Figure 3-7, the correlation peaks not when the point of first-order change (abrupt transition of a square wave) and the point of second-order change (the point of inflection of a triangular wave) are synchronized but at about 90 deg out of phase. This is because of a 90-degree phase shift of the fundamental frequency between square and triangular waves. Our finding that the inverse response increased for cross-waveform conditions supports the cross-correlation calculation idea.

In the cross-waveform conditions, the inverse response was the response of choosing a quadrants diagonally opposite the target. The inverse-response quadrant had the same waveform as the target while having a 90-degree phase shift from the target. The target's fundamental frequency phase is closer to the remaining non-target quadrants with a square-wave modulation. Note that two opposite phase modulations (e.g., 0-deg and 180-deg) were mixed in one region, so the phase shift of -90 deg and +90 deg did not matter much in the current argument. Our findings, therefore, are consistent with the hypothesis that the visual system compares the time courses of relatively raw sensory signals by computing the cross-area (dis)similarity through cross-correlation-like computation. However, there remains a higher degree of ambiguity in the algorithm details. On the other hand, these findings are inconsistent with the hypothesis that the visual system extracts universal temporal features (messengers) from stimulus changes in any order and then compares their timings across space to group or segment areas (Guttman et al., 2005).

Note that the cross-correlation computation does not contradict the previous findings about visual segmentation induced by timing misalignment (Kandil & Fahle, 2001; Sekuler & Bennett, 2001; for a review, Blake & Lee, 2005). The greater the timing misalignment (or the temporal phase shift) between the matched wave becomes, the lower the waveform similarity (e.g., cross-correlation) will be. Segmentation based on timing misalignment and that based on temporal waveform dissimilarity will predict similar results.

Nevertheless, simple cross-correlation computation does not explain why the asynchrony between second-order changes is less effective than between first-order changes. Even if the algorithm of computing temporal asynchrony adopted in the signal comparison stage (i.e., the second stage in Figure3-1) is shared across all the attributes and waveforms, they can affect the resolution of the temporal asynchrony detection by changing the temporal characteristics of the signals encoded in the first stage, and to be compared in the second stage. In addition, some mechanisms included in the comparison stage, such as temporal integration of the sensory signal, might make the asynchrony between triangular waves less salient than between square waves. In the current study, we cannot specify the precise source of the waveform effect, and we are currently developing detailed computational models of the segmentation process.

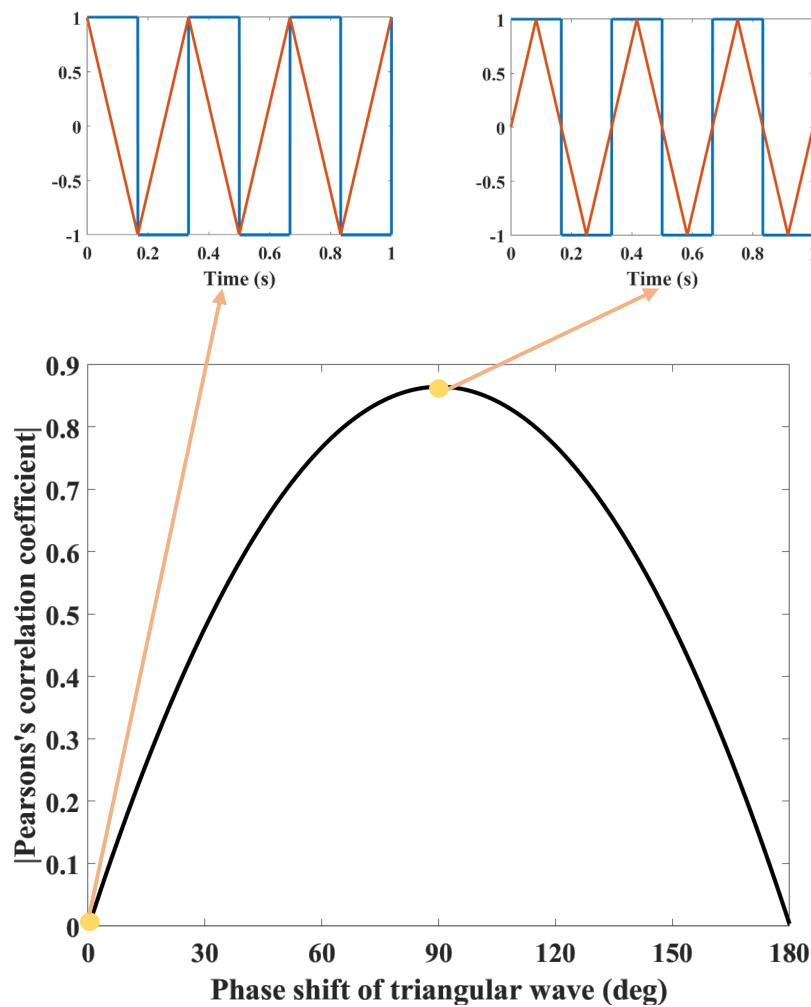


Figure 3-7. Pearson's correlation coefficient as the function of the phase shift of a triangular wave relative to a square wave.

The correlation between square and triangular waves is zero when the phase shift between the changing points is zero (i.e., the changing timings are aligned). In contrast, the correlation peaks when the phase shift between the changing points is 90 degrees (misaligned) since the fundamental frequencies are aligned.

3.4.2. Attribute-invariant process

Despite the general tendency of attribute invariance, some of our findings were incompatible with strict attribute invariance. First, there are minor differences among attributes even under the matched waveform conditions. For instance, we found lower performance at high temporal frequencies for color than for luminance. This difference may reflect how the early visual mechanisms respond to different attributes. Psychophysical research shows that the visual impulse response is more sluggish for color than for luminance (Burr & Morrone, 1993), which is at least qualitatively supported by

physiological findings (Mullen et al., 2010). Moreover, various stimulus factors, including contrast and spatial frequencies, influence the temporal response of the visual system (Watson, 1986). The variation of the tuning characteristics may at least partially be ascribed to the difference in the temporal characteristics (e.g., impulse response) of local attribute alternation detectors, the first stage in Figure3-1. Further investigation is needed to determine the extent to which changes in the impulse response function contribute to the observed variation in temporal tuning functions associated with asynchrony-based segregation.

The second finding incompatible with strict attribute invariance is that the performance was slightly worse for cross-attribute conditions than for within-attribute conditions, regardless of the attribute pairs. Explanation of this difference obviously requires a process beyond local attribute signal encoding. Some loss of performance for the cross-attribute may be ascribed to the mismatch in temporal response between different attributes to be compared in the second stage in Figure3-1. We may also need to consider additional noise factors associated with comparing and integrating different attributes across space (Meese, 2010).

Concerning the processing mechanism, our findings are consistent with the hypothesis that asynchrony detection for a variety of attribute combinations is processed by a single common comparison mechanism (Figure3-1). Critically, we found that tuning functions for different attributes were similar in shape as long as we matched the waveform. We also found that the tuning functions of the cross-attribute conditions were similar in shape to those of the corresponding within-attribute conditions and that the cross-attribute performance was never superior to the corresponding within-attribute conditions. Although a few deviations from strict attribute invariance were observed in current data, they do not contradict the common mechanism hypothesis, since they can be attributed to the early stages' tuning characteristics or mismatch of the compared signals.

However, functional attribute invariance in psychophysics does not necessarily imply a common underlying neural process. Our results are also consistent with an alternative hypothesis that the visual system has multiple comparison mechanisms specialized for specific attributes or for within- and cross-attributes, and all of them share similar processing characteristics. Further exploration is needed using both modeling and experimentation to draw a clear conclusion about the underlying mechanism.

3.4.3. Related Studies

Although studies have proposed that the timing of motion direction alternations could serve as an effective cue for asynchrony-based segmentation (Kandil & Fahle, 2004; Lee & Blake, 1999; Usher & Donnelly, 1998), at least under the conditions we tested, a direction reversal of continuous motion is found to be less effective than a luminance or color change as a cue for within-attribute asynchrony-based segregation. On the other hand, we demonstrate that an abrupt spatial phase shift, which induces an apparent motion, is as effective as a luminance or color change.

Our position differs from that of Farid & Adelson (2001), who claimed that it is not the motion direction reversal per se but rather the transient luminance contrast change produced by the direction reversal that acts as the temporal cue for segregation based on direction change asynchrony. Using their explanation, we expected that the segmentation would be easy when the motion direction change (P2) was paired with the luminance change (L1) since it is effectively an asynchronous detection of luminance changes. However, we found that the performance of the L1-P2 condition was low, close to the chance level. We accept that changes in the spatial phase or motion direction could be an effective cue for visual segmentation.

Lee & Blake (1999) proposed that differences in temporal structure between areas, rather than simple temporal asynchrony, are a strong segmentation cue (Blake & Lee, 2005). They showed that even when the alternation timing was occasionally aligned in a random sequence, participants still could perform the visual segmentation if the correlation of the two alternation timing series was low enough. These studies used random temporal alternations, whereas we used regular alternations. Motion direction changes at random intervals seem to be an effective segmentation cue. Why? One possibility is that differences in random temporal structure and asynchrony between repetitive changes drive different segmentation mechanisms, as Lee & Blake (1999) proposed. An alternative possibility is that they drive the same asynchrony-sensitive mechanism, but long intervals between each motion direction alternation occasionally appearing in the sequence may improve segmentation performance for random structure stimuli. Stated another way, since segregation is generally easier at low-temporal-frequency alternations, and a random sequence contains low-temporal-frequency modulation components, segmentation might be easier for random sequences than for rapid repetitive alternations. These issues should be investigated further.

3.4.4. Potential implementation of the attribute invariant computation on neurophysiology

Our conceptual framework for the underlying processing assumes a set of independent channels for luminance, color, and motion direction, potentially featuring significantly different temporal tunings for detecting alternations at the first stage. However, the underlying neural structures for detecting these attributes cannot be well specified, and they might exhibit some interaction and overlap. For instance, neurophysiological studies have highlighted that tuning characteristics of chromatic and achromatic stimuli may not be markedly different in the primary stages of visual information processing, such as LGN or V1 (D'Souza et al., 2011; Mullen et al., 2010; for a book chapter, Kremers et al., 2016), as compared to those indicated by psychophysical performances. There is also evidence showing that the neurons in V4 may also show direction selectivity, suggesting the overlapping mechanism for color and motion (Tolias et al., 2005). Notably, the activity of a set of neurons does not necessarily explain perceptual experience. As Jiang et al. (2007) demonstrated, a color patch flickering at sufficiently high frequency can be perceived as static, yet the response in V1 to V4 cortex persists, indicating a dissociation between subjective experience and neuron response in the early cortex response. Hence, we cannot conclude which set of neurons corresponds to the assumed first stage. The actual neural architecture may be considerably more complex than we currently assume.

Regarding the asynchrony comparator in the second stage, there is also no consensus about which area in the cortex is responsible for determining temporal relationships. Also, the area might be distributed over the cortex. It is suggested that the Inferior Parietal Lobe (IPL) in the right hemisphere deficiency impairs the out-of-phase detection (Battelli et al., 2003), while the insula is activated for visual-audio asynchrony detection (Bushara et al., 2001).

We do not necessarily assume direct correspondence between the underlying neural architecture and the model shown in Fig. 1. However, even though the channels or architectures can be highly complicated and interactive, we expect the algorithm of neural signal processing to be consistent, particularly in extracting the amodal signal for temporal relationship comparison and asynchrony signal detection. Recent MRI studies, such as those conducted by Groen et al. (2022) and Zhou et al. (2018), reveal analogous linear-nonlinear structures representing temporal information from V1 to V3.

Additionally, correlation calculation is frequently considered a candidate for multisensory asynchrony detection algorithms, as observed in studies predicting EEG or MRI data (Parise & Ernst, 2016; Pesnot-Lerousseau et al., 2022). The recent study also points out that spatiotemporal correlation can be a computation for the system conducting the spatial pooling or integration (Mehrani & Tsotsos, 2023; Sun et al., 2023a, 2023b)

3.5. Conclusion

Existing studies partially support an attribute-invariant process that relies primarily on temporal information for the visual segmentation task. In our experiments, we assessed segmentation performance across a more diverse attribute setup and demonstrated the prominence of an attribute-invariant computation when waveform matching is ensured. The observed impact of the waveform challenges the previously posited hypothesis that the simple timing of changes, irrespective of temporal order and attribute, can be compared directly across space. Our discoveries prompt a re-evaluation of the hypothesis. A deeper exploration into how simultaneity is determined across space is essential to uncover the mechanism underlying spatiotemporal visual segmentation.

Chapter 4. Single Process Channel Underlying Temporal Correlation Based Visual Segmentation

In the previous chapter, we identified the presence of a feature-invariant process underlying temporal segmentation, and suggested that the visual system may be sensitive to temporal correlation across different attributes. In this chapter, we further investigate the origin of this feature invariance. Specifically, we examine whether it arises from a single, unified processing mechanism that handles all attributes collectively, or from multiple attribute-specific channels that converge at a later stage through a common integration process.

4.1. Introduction

4.1.1. The feature integration issue in temporal segmentation

Visual segmentation is a fundamental function of the visual system that allows us to partition the visual space into distinct objects. A widely accepted computational principle for segmentation is to assess the similarity or dissimilarity of elements across a boundary (James R. Bergen & Landy, 1991; Michael S. Landy & Graham, 2003). When elements are similar, they are perceived as part of the same object or texture; otherwise, they are perceived as separate entities. The visual system evaluates similarity across multiple attributes, including luminance, orientation, spatial frequency, and chromatic hue (James R. Bergen & Landy, 1991; Callaghan et al., 1986; Dakin & Watt, 1997; H. C. Nothdurft, 1985; H.-C. Nothdurft, 2007).

While many studies have explored segmentation based on a single attribute, natural visual scenes often contain objects defined by multiple attributes. This raises the question of how the visual system integrates different attributes to form a coherent percept of grouping or segmentation. Prior research has demonstrated that the system effectively integrates different visual cues, such as luminance contrast and orientation (M. S. Landy & Kojima, 2001), luminance texture and chromatic hue (Saarela & Landy, 2012), or the spatial frequency and orientation (Hunt & Meinhardt, 2021). Moreover, studies suggest that this integration is mandatory—it occurs automatically and cannot

be selectively ignored, implying the presence of feature-invariant mechanisms (Saarela & Landy, 2012, 2015).

While feature integration in static scenes has been well studied, everyday vision is inherently dynamic. Thus, it is crucial to ask whether similar integration principles apply when features vary over time.

4.1.2. Temporal correlation as an attribute-invariant processing

Grouping and segmentation in dynamic contexts often rely on temporal correlation—the synchronized variation of visual attributes. When attributes covary over time, the corresponding elements are perceived as grouped; otherwise, they are perceived as separate. This principle underlies various perceptual phenomena, including common fate, where motion direction and speed change synchronously (Uttal et al., 2000; Wertheimer, 1923), as well as the temporal covariation of luminance, (Alais et al., 1998; Sekuler & Bennett, 2001; Usher & Donnelly, 1998), spatial frequency and orientation (Guttman et al., 2005), and color (Chen et al., 2024a). A more comprehensive discussion of these effects can be found in Blake & Lee (2005).

Interestingly, research suggests that observers can simultaneously integrate multiple types of covarying features when performing grouping tasks (Chen et al., 2024a; Guttman et al., 2005). This raises a fundamental question: How does the visual system integrate different temporally correlated attributes to influence segmentation? While prior findings indicate that temporal correlation acts as a general computational mechanism independent of specific features (Chen et al., 2024a, 2025; Nishida & Johnston, 2002), the precise way in which different temporally correlated attributes interact remains unclear.

4.1.3. Single or multiple channels underlying the process?

Building on these insights, our study aims to investigate how temporal correlations across different attributes interact to shape visual segmentation. Specifically, we seek to determine whether the integration of different attributes occurs before or after calculating temporal correlation (see Figure 3-8 for a demonstration). If integration occurs before calculating temporal correlation, signals from different attributes must first be

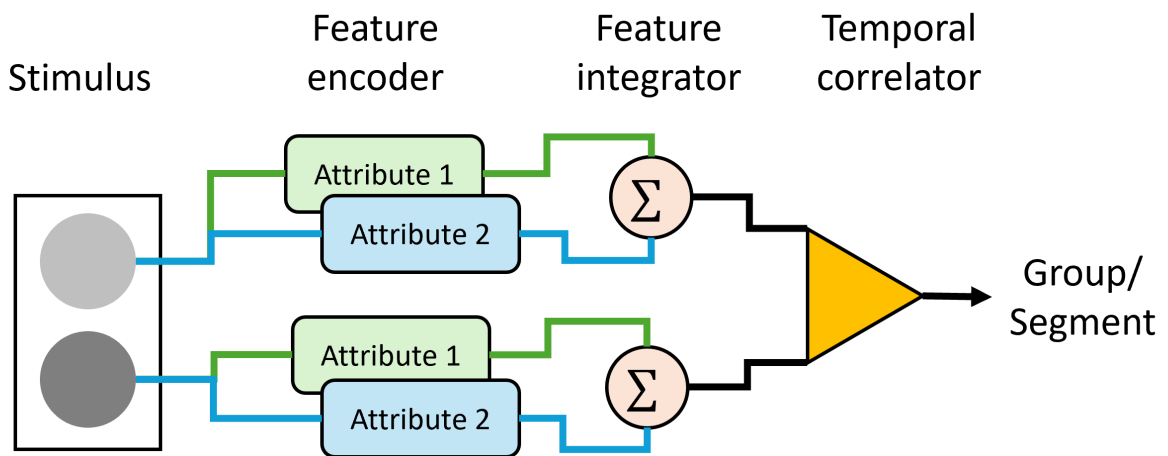
combined into a unified representation before the correlation is computed (case 1 in Figure 3-8). This would imply a single computation channel, where different attributes influence each other, ultimately modulating the perceived correlation magnitude. In contrast, if integration occurs after calculating temporal correlation, each attribute's correlation is computed separately before they are combined at a later stage (case 2 in Figure 3-8). This would suggest the presence of multiple independent computation channels, where the correlation magnitude remains unaffected by cross-attribute interactions.

To address this question, we implemented a noise-masking paradigm to evaluate segmentation performance using stimuli developed by Chen et al., (2025). These stimuli systematically varied temporal correlation levels while preserving broad-spectrum characteristics, ensuring that all potential channels in low-level vision were adequately activated. In our experiment, we measured segmentation performance under two conditions: (1) within-attribute noise, where noise was added to the same attribute channel as the signal, and (2) cross-attribute noise, where noise was introduced in a different attribute channel. We then compared segmentation performance in the cross-attribute condition relative to the within-attribute condition. If segmentation performance in the cross-attribute condition is similar to that in the within-attribute condition, this would suggest that noise from a different attribute channel interferes with segmentation to a comparable degree as noise within the same attribute channel. This outcome would support the single-channel hypothesis, where attribute integration occurs before temporal correlation is computed, allowing cross-attribute interactions to influence segmentation. Conversely, if segmentation performance in the cross-attribute condition remains significantly higher than in the within-attribute condition—approaching the level observed in the absence of noise—this would indicate that noise from different attributes does not interfere with the signal. Such a result would support the multiple independent-channel hypothesis, where temporal correlation is computed separately for each attribute before integration, preventing cross-attribute interference.

In this research, we carried out two experiments and created a computational model. Experiment 2-1 focused on directly assessing how segmentation performance changed with temporal correlations under conditions of noise within a single attribute or across different attributes, specifically examining three attributes: luminance, color (red/green), and spatial phase. Experiment 2-2 explored varying levels of local

luminance variation to determine whether and how the strength of low-level signals affects the extent of noise interference. Ultimately, the computational model offers a coherent framework to interpret our experimental results.

Case 1. Single channel: integrate **before** correlation calculation



Case 2. independent channel: integrate **after** correlation calculation

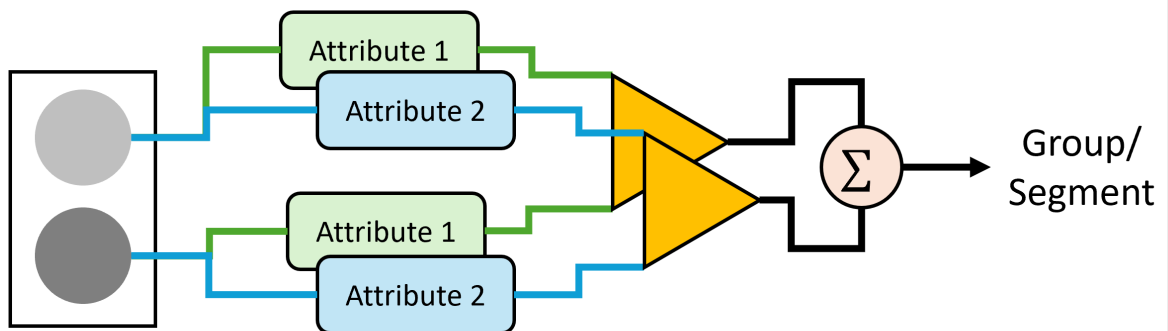


Figure 3-8. Two putative process for the temporal correlation-based visual segmentation.

4.2. Experiment 2-1. The Correlation Noise Interference from Same and Different Visual Attributes

4.2.1. Methodology

4.2.1.1. Apparatus

The stimuli were displayed on a VIEWPixx /3D LCD monitor (VPixx Technologies, Saint-Bruno-de-Montarville, Canada) with a 1920×1080 pixels resolution and set as refreshing the screen of 30 Hz. The lowest, mean, and highest display luminance was 1.8, 48.4, and 96.7 cd/m^2 , respectively. The linear output of the luminance was calibrated using an i1Pro chromometer (VPixx Technologies) for each channel. Each pixel subtended 1.3 arcmin at a viewing distance of 70 cm. The participant was seated in a dark room with a chinrest stabilizing the head. The experiments were programmed using the Psychophysics Toolbox (Brainard, 1997) in MATLAB release 2022a (Math Works, Natick, MA, USA). The study protocol followed the ethical standards of the Declaration of Helsinki except for preregistration and was approved by The Ethics Committee of Kyoto University (KUIS-EAR-2020-003).

4.2.1.2. Participants

The experiment included two of the authors and three naïve participants (five males, mean age 24.8 years). The participants declared that they had normal or corrected-to-normal vision. All the participants were informed of the nature of the research and provided informed consent before the experiment started. Participants received monetary compensation after the experiment finished. A revised version of the minimum motion method (Anstis & Cavanagh, 1983; Cavanagh et al., 1987) was used to measure the equiluminant ratio (in terms of the system intensity value) between the red (R) and green (G) channels for each participant before the main experiment. During the procedure, participants saw a Gabor drifting rightward. The sinusoidal wave of the R and G channels is 180 degrees out-of-phase. Participants were required to adjust the ratio between the channels to minimize the sense of motion. The mean R: G ratio value was 0.51.

4.2.1.3. Stimuli

4.2.1.4. Redefining the feature tensor

We utilized the feature tensor generated by Chen et al. (2025), where each tensor consists of an 8×16 spatial grid spanning 30 frames. The values within the tensor range from -1 to 1. This tensor can potentially be divided into left and right partitions based on the temporal correlation within each section. We denote the generated tensor as $F(i, j, t; \phi)$, where (i, j) represents the spatial location, t denotes the time frame, and ϕ specifies the temporal correlation magnitude. For each correlation level, 30 tensors were generated, covering a total of 11 correlation levels ranging from 0 to 1.

To systematically examine segmentation performance, we introduced variations in attribute combinations and spatial arrangements. The attribute combinations include two conditions: (1) within-attribute condition, where the stimuli only composed of one of the three visual attribute, and (2) cross-attribute condition, where the stimuli composed of two of the three visual attribute. The spatial arrangements are categorized into three types: (1) full-signal, where all elements deliver the signal, (2) half-empty, where only half of the elements are visible, and (3) half-noise, where half of the elements are replaced with random noise. It is important to note that the half-empty condition applies only to the within-attribute condition and is not present in the cross-attribute condition.

In the within-attribute condition, the revised feature tensor is defined as follows:

$$\begin{aligned}
 & F'(t; \phi) \\
 &= \begin{cases} F(i, j, t; \phi), & \text{full - signal} \\ M(i, j) \odot F(i, j, t; \phi), & \text{half - empty} \\ M(i, j) \odot F(i, j, t; \phi) + (1 - M(i, j)) \odot \epsilon(i, j, t), & \text{half - noise} \end{cases} \quad (3-14) \\
 & M(i, j) \sim \text{Bernoulli}(p = 0.5) \\
 & \epsilon(i, j, t) \sim \text{Unif}(-1, 1)
 \end{aligned}$$

Where $M(i, j)$ is a masking function that randomly determines which elements are displayed, and $\epsilon(i, j, t)$ is a random noise draw fromd uniform distribution from -1 to 1.

In the cross-attribute condition, the feature tensor is modified as follows:

$$\begin{aligned}
& F'(t; \phi) \\
&= \begin{cases} A(i, j) \odot F(i, j, t; \phi) + (1 - A(i, j)) \odot F(i, j, t; \phi), & \text{full - signal} \\ A(i, j) \odot F(i, j, t; \phi) + (1 - A(i, j)) \odot \epsilon(i, j, t), & \text{half - noise} \end{cases} \quad (3-15)
\end{aligned}$$

$$A(i, j) \sim \text{Bernoulli}(p = 0.5)$$

$$\epsilon(x, y, t) \sim \text{Unif}(-1, 1)$$

Here, $A(i, j)$ replace $M(i, j)$ determines whether a location displays the first or second attribute. Therefore, notice that one attribute can only deliver a signal or noise during the half-noise situation.

4.2.1.5. Actual stimuli generation

The final stimulus is constructed based on the revised feature tensor $F'(i, j, t; \phi)$, randomly selected from the 30 generated tensors. The stimulus consists of an 8×16 dynamic lattice, presented for 30 frames (1 second). Each element subtends 0.5° in diameter, forming an overall lattice size of $4^\circ \times 8^\circ$. The values in F' are mapped to individual elements within the lattice, where each element corresponds to one of three attributes—luminance, color, or spatial phase—depending on the attribute condition and the mask $A(i, j)$. For within-attribute conditions, all elements are assigned to a single attribute (e.g., all elements defined by luminance). In cross-attribute conditions, such as the luminance-color combination, an element is defined by luminance when $A(i, j) = 1$ and by color otherwise, other attribute pair also follows the same rule.

When the attribute is luminance, the element is defined as a Gaussian bulb with following form:

$$L(x, y, t) = L_{mean} \cdot \left(1 + F'(i, j, t) \cdot e^{\frac{x^2 + y^2}{\sigma^2}} \right) \quad (3-16)$$

x and y are the spatial locations on the plane, where 0 is the center of the element, and t is the frame. i, j is the location within the F' , where i is integer among 1 to 8, and j is integer among 1 to 16. L_{mean} is the mean luminance of the screen, σ is the Gaussian windows of the element, which is set as 5.25 arcmin.

When the attribute is color, the element also is a Gaussian bulb but with three separate channel assignments:

$$L(x, y, t) = \begin{cases} R(x, y, t) = L_{mean} \cdot \left(1 + F'(i, j, t) \cdot e^{\frac{x^2+y^2}{\sigma^2}} \right) \\ G(x, y, t) = L_{Iso.RG} \cdot L_{mean} \cdot \left(1 - F'(i, j, t) \cdot e^{\frac{x^2+y^2}{\sigma^2}} \right) \\ B(x, y, t) = L_{mean} \end{cases} \quad (3-17)$$

$L_{Iso.RG}$ is the iso-luminant index between red and green.

When the attribute is spatial phase, the element is changed as the Gabor patch with following form:

$$L(x, y, t) = L_{mean} \cdot \left(1 + \cos \left(2\pi\omega r - \frac{\pi}{4} F'(i, j, t) \right) \cdot e^{\frac{x^2+y^2}{\sigma^2}} \right) \quad (3-18)$$

$$r = y\cos(\theta) + x\sin(\theta)$$

ω is the spatial frequency, which is set as 2 cpd, θ is the orientation of the Gabor pattern, which is randomly drawn from the uniform distribution amongst $[0, 2\pi)$. The $\frac{\pi}{4}$ in advance the F^* make sure the spatial phase change would not jump surpass 90 degree per frame to prevent the effects of the luminance flickering.

4.2.1.6. Design and procedure

To assess segmentation performance, we employed a two-interval forced-choice (2-IFC) task. Each trial consisted of two consecutive 1-second video clips—one containing a stimulus that could be perceptually segmented into left and right partitions, while the other remained unsegmented. The sequence of these two clips was randomized across trials. After both clips were presented, participants were required to indicate which one contained the segmented structure by pressing a designated key: the left arrow key for the first clip and the right arrow key for the second clip. Each task condition included 11 levels of temporal correlation, ranging from 0 to 1 in increments of 0.1. Temporal correlation was defined as the absolute correlation of temporal cosine similarity within the left or right partition, while the correlation between the left and right partitions was fixed at 0. This ensured that all correlations greater than zero had the potential to be perceptually segmented. The sequence of trials with different correlation levels was randomized within each condition. As described in the stimulus section, the experiment included two types of attribute combinations (within-attribute and cross-attribute) and multiple spatial arrangement conditions. For the within-attribute condition, three spatial arrangements were tested: full-signal, half-empty, and half-noise. For the cross-attribute

condition, only full-signal and half-noise arrangements were included. Additionally, the cross-attribute condition encompassed all possible attribute pairings: luminance-color (Lum-Col), luminance-spatial phase (Lum-Pha), and color-spatial phase (Col-Pha). In cross-attribute half-noise conditions, only one of the attributes served as the signal in each trial, and the designated signal attribute was randomized across trials. The sequence of condition pairings was also randomized. In total, the participants need to finish trial number $(30) \times \text{correlation}(11) \times \text{spatial arrangement}(3) \times \text{attribute type}(3) = 2970$ trials for within-attribute condition, and trial number $(30) \times \text{correlation}(11) \times \text{spatial arrangement}(2) \times \text{attribute pair}(3) = 1980$ trials for cross-attribute condition. The experiment took approximately 6 hours to complete.

4.2.1.7. Analysis

To quantify how closely the cross-attribute condition performance aligns with the respective hypotheses (i.e., single or multiple correlation computation channels), we compare the relative distance of cross-attribute noise to the within-attribute half-empty and half-noise conditions, respectively. The distance to half-empty is used instead of full-signal because the number of elements delivering the signal in the half-empty condition is equal to that in the cross-attribute half-noise condition, whereas in the full-signal condition, the number is doubled. Furthermore, the relative distance to the two hypotheses is indexed by the ratio of the relative distance to half-empty to the relative distance to half-noise. The formula is as follows:

$$DH = \log_{10} \left(\frac{\sum_{\phi} (R_{CAHN}(\phi) - R_{WAHE}(\phi))^2}{\sum_{\phi} (R_{CAHN}(\phi) - R_{WAHN}(\phi))^2} \right) \quad (3-19)$$

where DH represents the distance to hypothesis, with various subscripts indicating the proportion of participants who responded correctly under different conditions. Specifically: CAHN refers to cross-attribute half-noise, WAHE denotes within-attribute half-empty, and WAHN signifies within-attribute half-noise. Thus, the numerator represents the Euclidean distance between cross-attribute half-noise and within-attribute half-empty, while the denominator represents the Euclidean distance between cross-attribute half-noise and within-attribute half-noise.

- If $DH > 0$, it indicates that the numerator is larger than the denominator, meaning that the distance from CAHN to WAHE is greater than the distance from

CAHN to WAHN. This suggests that cross-attribute noise significantly affects performance, aligning more with the single-channel hypothesis.

- Conversely, if $DH < 0$, it implies that the numerator is smaller than the denominator, meaning that the distance from CAHN to WAHE is less than the distance to WAHN. This suggests that cross-attribute noise does not significantly affect performance, supporting the independent-channel hypothesis.

To estimate the error of each data point, we adopt a bootstrapping procedure. Each data point represents the average proportion correct across five participants. In each iteration of the bootstrapping process, we resample five values from the original five participants' data with replacement. This procedure is repeated for a total of 2,500 iterations. The 2.5th and 97.5th percentiles of the bootstrapped distribution are used to determine the confidence interval for each data point.

4.2.2. Result

4.2.2.1. Noise disturbance within same attribute channel

Figure 3-9 shows proportion correct as a function of temporal correlation strength for each visual attribute—luminance, color, and spatial phase—under the within-attribute condition. Across all attributes, segmentation performance increased with increasing correlation, indicating that temporal correlation reliably supports visual segmentation. In the full-signal condition (black lines), participants reached near-ceiling performance when correlation was high. In the half-empty condition (teal), where only half the elements delivered signal, performance followed a similar trend but remained slightly lower than full-signal condition across the correlation range. In contrast, the half-noise condition (orange) consistently yielded the lowest performance, particularly at mid-to-high correlation levels. This pattern was observed across all three attributes. The performance gap between half-empty and half-noise conditions demonstrates that the addition of temporal noise in the same attribute channel substantially impairs segmentation beyond the effect of reduced signal density alone.

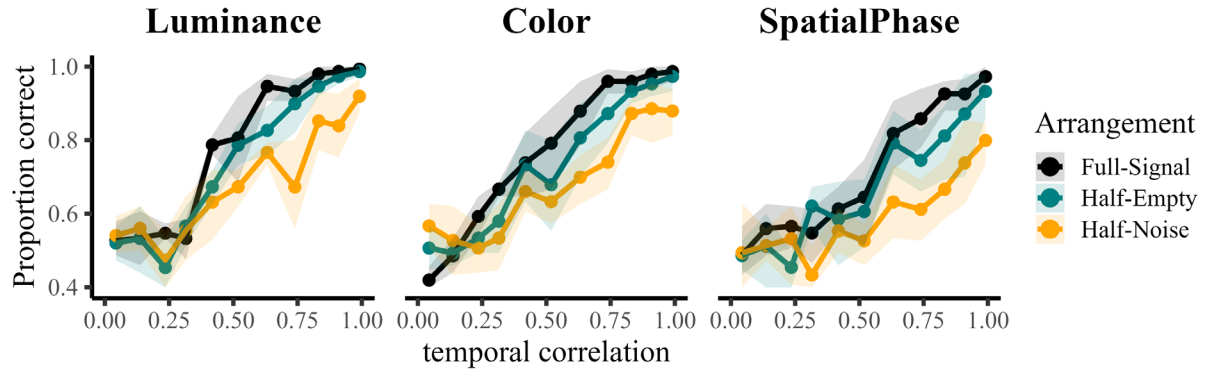


Figure 3-9. Segmentation performance as a function of temporal correlation strength in the within-attribute condition.

Each panel corresponds to a different visual attribute: luminance (left), color (middle), and spatial phase (right). The three spatial arrangements include full-signal (black), half-empty (teal), and half-noise (orange) conditions. Shaded areas represent the 95% confidence intervals, computed from the 2.5th to 97.5th percentile of the bootstrapped distribution across participants.

4.2.2.2. Noise disturbance across attribute channels

Figure 3-10 shows segmentation performance in the cross-attribute condition across all attribute combinations. For the convenience of discussing, we abbreviated all the attribute pair as follows: Lum-Col for luminance and color; Lum-Pha for luminance and spatial phase; Col-Pha for color and spatial phase. In each panel, black lines represent the full-signal condition meaning both attribute contain the signal is capable to do the segmentation, and red lines represent the cross-attribute half-noise condition. For reference, performance in the within-attribute half-empty (teal) and half-noise (orange) conditions is also plotted.

In general, segmentation performance in the cross-attribute half-noise condition (red) was equal to or lower than that in the cross-attribute full-signal condition (black), with the extent of interference depending on which attribute served as the signal. When luminance delivered the signal (first column in Figure 3-10), color and spatial phase noise produced minimal interference, and performance remained largely unaffected. In contrast, when color or spatial phase was the signal (first row in Figure 3-10), luminance noise substantially impaired segmentation performance. In contrast, the color and spatial phase mutually interfered in a similar extent.

To further examine whether cross-attribute noise produces interference comparable to within-attribute noise, we computed the DH index, which quantifies the relative proximity of cross-attribute performance to the within-attribute half-noise and half-

empty baselines (Figure 3-11). The results show that luminance, when introduced as noise, produces a level of interference on color and spatial phase signals comparable to that observed in the respective within-attribute noise conditions. In contrast, color and spatial phase noise had little effect when luminance served as the signal. Additionally, spatial phase noise tended to produce stronger interference than color when acting as the noise source.

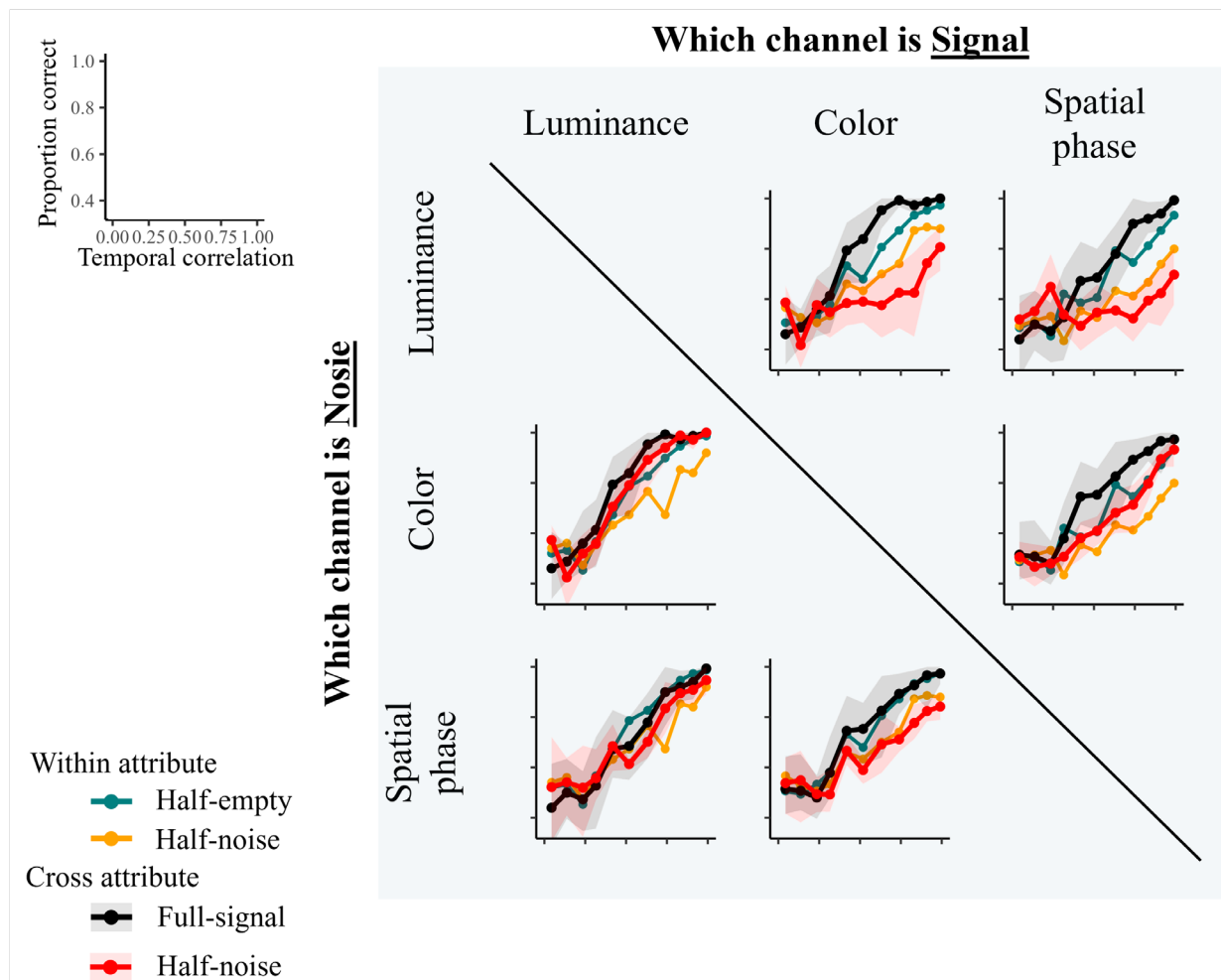


Figure 3-10. Segmentation performance in the cross-attribute condition across all combinations of signal and noise attributes.

Each panel corresponds to a unique pairing, with columns indicating the signal attribute (from left to right: luminance, color, spatial phase) and rows indicating the noise attribute (from top to bottom: luminance, color, spatial phase). Two conditions are plotted for each pair: cross-attribute full-signal (black) and cross-attribute half-noise (red). For comparison, performance in the within-attribute half-empty (teal) and within-attribute half-noise (orange) conditions is also shown as reference baselines. Shaded areas represent the 95% confidence intervals, computed from the 2.5th to 97.5th percentile of the bootstrapped distribution across participants

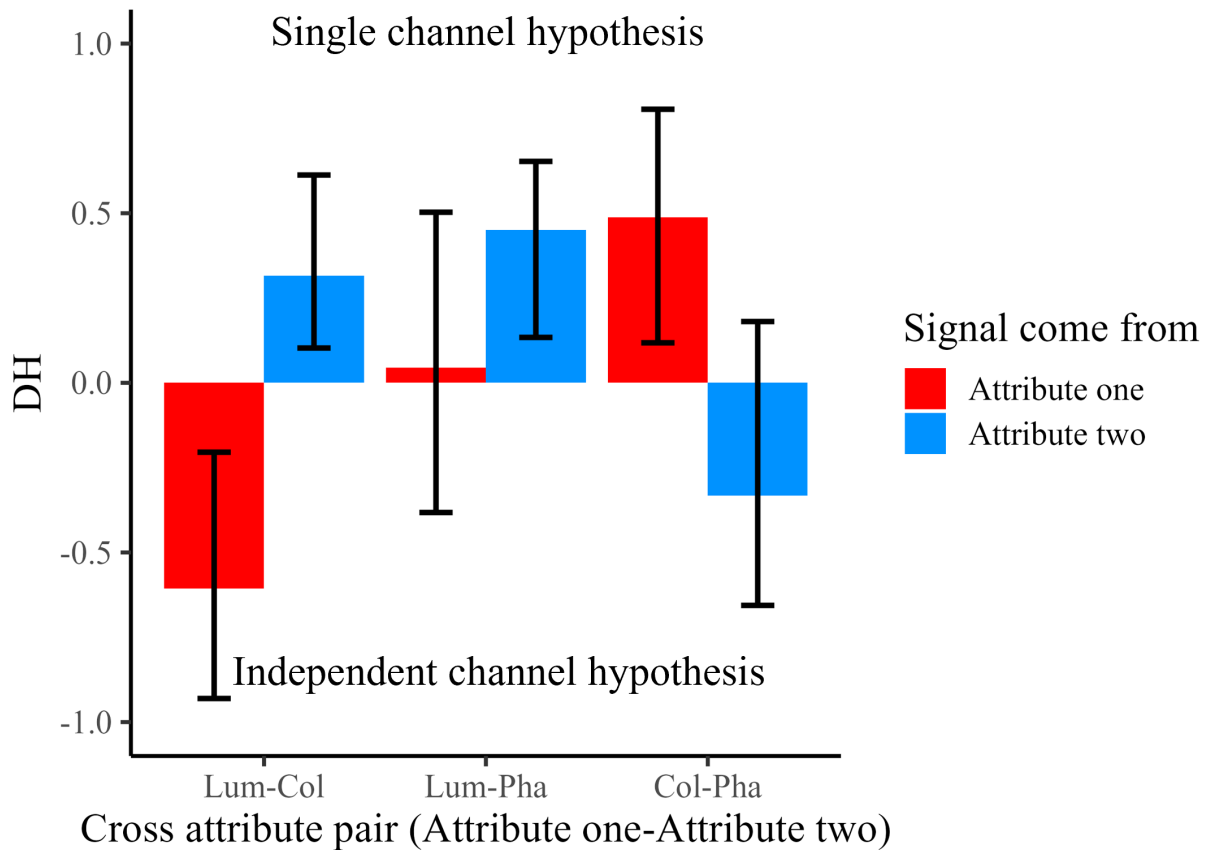


Figure 3-11. Distance index DH for all cross-attribute pairings

DH quantifies the relative position of cross-attribute performance between the within-attribute half-empty and half-noise baselines. Bars represent the mean DH value for each attribute pair, labeled in the format “Attribute1–Attribute2”. Red bars correspond to conditions where the signal comes from the first attribute in the label, and blue bars indicate conditions where the signal comes from the second attribute. Positive values indicate that cross-attribute performance is closer to the within-attribute half-noise condition, while negative values reflect proximity to the half-empty baseline. Error bars represent the 95% confidence intervals, computed from the 2.5th to 97.5th percentiles of the bootstrapped participant distribution.

4.2.3. Interim discussion

In Experiment 2-1, we measured segmentation performance under within-attribute and cross-attribute conditions to examine whether noise introduced in one feature channel could interfere with segmentation based on another channel—similar to how noise disrupts performance within the same attribute. The results suggest that cross-channel noise can plausibly interfere with signal processing in other channels and impair segmentation performance. In some cases, the disruptive effect of cross-attribute noise reached a level comparable to that observed in the within-attribute condition. Notably, the magnitude of this interference depended on the specific attribute pairing, with luminance acting as a particularly strong source of disruption when introduced as noise.

Since cross-attribute noise can interfere with segmentation performance, it is difficult to fully support the independent-channels hypothesis, which assumes that correlation computations occur entirely separately for each feature dimension. At the same time, the results do not provide sufficient support for the single-channel hypothesis either, as segmentation based on luminance signals remained unaffected by noise from color or spatial phase.

If the single-channel hypothesis were valid, one possible explanation for the robustness of luminance could lie in its relatively stronger internal representation compared to the other attributes. In this case, even if signals are merged prior to correlation computation, the dominant strength of luminance may overshadow any interfering noise. Accordingly, Experiment 2-2 was designed to test whether reducing the strength of luminance signals would make them more vulnerable to interference from color and spatial phase noise. This approach allows us to test whether the lack of interference observed in Experiment A was due to a genuinely independent processing channel for luminance, or simply a result of the signal strength imbalance between attributes.

4.3. Experiment 2-2. The Influence of Local Variation Magnitude

Experiment 2-1 revealed that while cross-attribute noise can impair segmentation performance, luminance-based segmentation remained largely unaffected by noise from color or spatial phase. One possible explanation is that luminance signals, due to their relatively stronger input strength, dominate the computation and override interference from other attributes—even if signals are integrated prior to the computation of temporal correlation. To test this possibility, Experiment 2-2 examined whether reducing the strength of luminance signals would increase their susceptibility to cross-attribute interference. The stimulus generation and task procedure were identical to those in Experiment 2-1, except that the luminance amplitude was scaled to 10%, 40%, or 70% of the full variation used previously. This manipulation was implemented by multiplying the original luminance signal by a scalar factor, while all other parameters were held constant. Only the cross-attribute half-noise condition was tested. Participants from Experiment 2-1 also participated in Experiment 2-2, completing trials across 3 luminance variation levels $\times$ 2 signal attributes (color or spatial phase) $\times$ 11 correlation levels, with 30 repetitions per condition. This resulted in a total of 1,980 additional trials per participant, requiring approximately two hours to complete.

4.3.1. Result

Figure 3-12 showed the segmentation performance in the cross-attribute half-noise condition across different levels of luminance variation magnitude (100%, 70%, 40%, 10%), with the attribute pair is Lum-Col (top row) or Lum-Pha (bottom row). For both color and spatial phase as the signal (i.e., luminance is the noise), the performance is increasing with the decreasing luminance variation magnitude. Especially we observed a reversed effect where when the magnitude is 40 %, the segmentation performance based on the luminance is consistently surpass both color and spatial phase while when the magnitude is 10 %, the performance based on color and spatial phase is consistently surpass the luminance one. This suggests that the interference effect from the luminance is gradually decreasing since a lowering variation magnitude.

To quantify this change in interference, we used the same DH index as in Experiment 2-1, as shown in Figure 3-13. In the luminance–color condition (left panel), DH values for color as the signal (blue) remained consistently positive at higher luminance magnitudes, but dropped markedly at 10%, indicating reduced interference from luminance noise. A comparable pattern was observed in the luminance–spatial phase condition (right panel), where DH values for spatial phase as the signal (green) similarly declined with decreasing luminance strength. In contrast, when luminance was the signal, DH values remained at or below zero for luminance magnitudes of 40% and above, indicating that color and spatial phase noise did not impair segmentation performance. However, when luminance variation was reduced to 10%, the DH value increased above zero, suggesting that color and spatial phase noise began to interfere with luminance-based segmentation at a level comparable to within-attribute luminance noise.

These results confirm that the magnitude of luminance variation directly modulates its ability to interfere with other attributes. When luminance signals were strong, they caused substantial cross-attribute disruption; when weakened, their interference decreased to levels consistent with the half-empty baseline. This supports the interpretation that the asymmetry observed in Experiment 2-1 was not due to fully independent processing across feature channels, but rather to differences in relative signal strength among attributes.

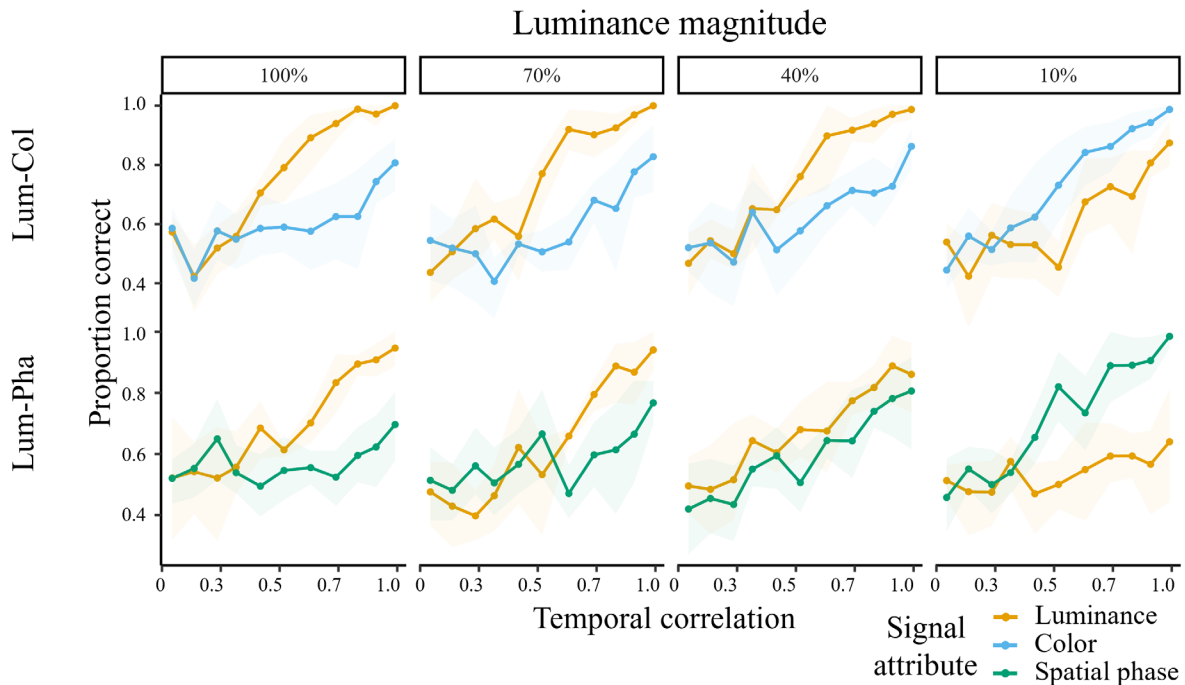


Figure 3-12. Segmentation performance in the cross-attribute half-noise condition of Experiment 2-2.

The result, shown as a function of temporal correlation for each luminance magnitude level (100%, 70%, 40%, 10%). Columns represent different levels of luminance variation. Rows represent attribute pairings, with the first row corresponding to the luminance–color pair and the second row to the luminance–spatial phase pair. Line colors indicate the signal attribute: orange for luminance, blue for color, and green for spatial phase. Shaded areas represent the 95% confidence intervals, computed from the 2.5th to 97.5th percentiles of the bootstrapped participant distribution.

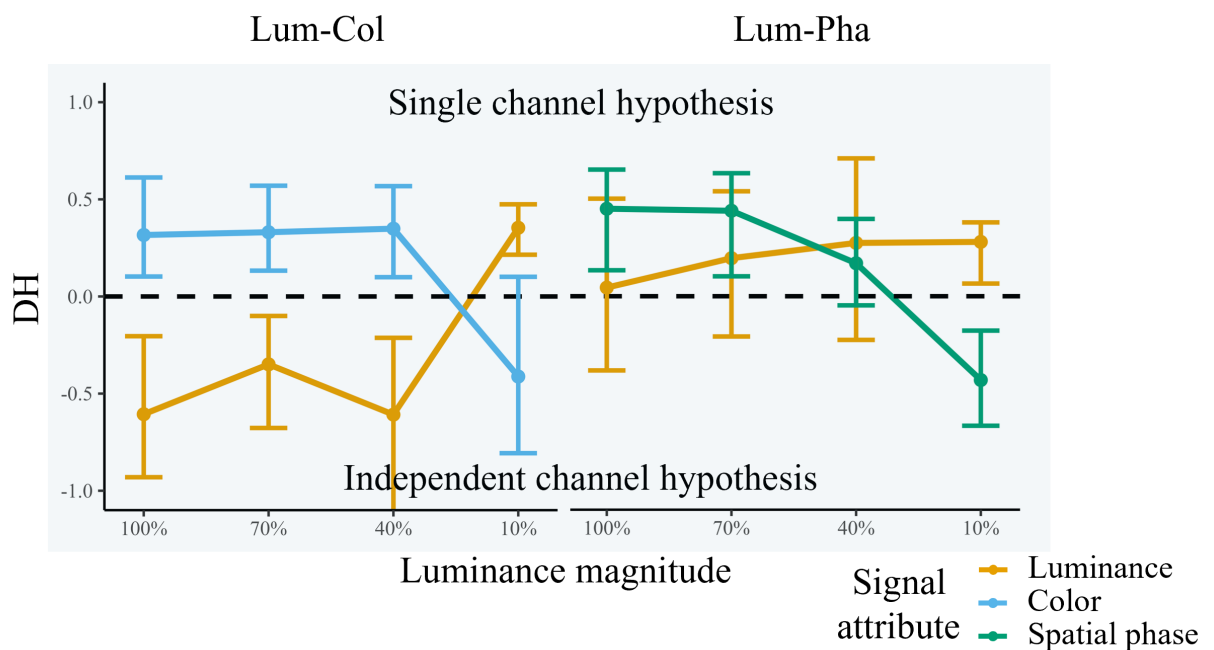


Figure 3-13. Distance index DH as a function of luminance magnitude in Experiment 2-2.

The result, shown separately for two attribute pairs: luminance–color (Lum–Col, left panel) and luminance–spatial phase (Lum–Pha, right panel). The DH index quantifies the relative position of cross-attribute segmentation performance between the within-attribute half-empty and half-noise baselines. Positive values indicate performance closer to the half-noise condition (consistent with the single-channel hypothesis), while negative values indicate proximity to the half-empty condition (consistent with the independent-channel hypothesis). Line colors indicate the signal attribute: orange for luminance, blue for color, and green for spatial phase. Error bars represent the 95% confidence intervals, computed from the 2.5th to 97.5th percentiles of the bootstrapped participant distribution.

4.4. Model Conduction – A Single Channel Hypothesis

The results from Experiments 2-1 and 2-2 suggest that segmentation based on temporal correlation is not computed in entirely separate channels for each visual attribute. Cross-attribute noise can impair performance, and the extent of interference depends on the relative strength of signals across attributes. These findings are inconsistent with a strictly independent-channel account but can be explained under a single-channel framework, in which signals from different attributes are integrated before correlation is computed. To further evaluate this possibility, we sought to formalize the single-channel hypothesis into a computational model. Our goal is to determine whether a single integration process—modulated by relative signal strength—can account for both the overall segmentation performance and the asymmetric interference patterns observed across attribute pairings. In particular, we ask whether differences in

interference strength can be explained by simple differences in gain across channels, without invoking multiple independent correlation mechanisms.

4.4.1. Model description

To describe segmentation performance as a function of temporal correlation, we adopted a Weibull-like psychometric function to fit the behavioral data, defined as:

$$P(\phi) = 1 - 0.5e^{-\beta(\Theta)\phi^2} \quad (3-20)$$

Here, 0.5 reflects the chance level of a two-interval forced-choice (2IFC) task. $\beta(\Theta)$ determines the steepness of the psychometric function and is governed by a parameter vector Θ , which is shared across all task conditions. The slope $\beta(\Theta)$ is modeled as a monotonic function of the signal-to-noise ratio (SNR) between the external signal and the combined internal and external noise:

$$\beta(\Theta) = A \cdot (1 - e^{-SNR}) \quad (3-21)$$

$$SNR = \frac{S}{\sigma_{inner} + N}$$

Here, A is a global scaling factor controlling the overall slope of the function, and σ_{inner} represents internal system noise. The signal (S) and noise (N) are computed as gain-weighted sums of the corresponding attribute vectors. The gain for luminance is fixed at 1, while G_{Col} and G_{pha} are free parameters for the gain of color and spatial phase respectively:

$$S = N = 0.5 \cdot \begin{bmatrix} 1 \\ G_{Col} \\ G_{pha} \end{bmatrix}^T \cdot (I_1 + I_2) \quad (3-22)$$

Each I_i is a 3-element vector indicating the normalized strength of signal or noise in the luminance, color, and spatial phase channels, respectively. Two different vectors reflect the experimental design in which each of the 8×16 stimulus elements was randomly assigned to one of two spatially intermixed subsets (with equal probability), and each subset was independently manipulated to carry signal or noise. As such, I_1 and I_2 represent the average attribute composition in these two independently controlled subgroups.

To illustrate how signal and noise vectors are defined under specific conditions, several examples are provided below:

- Within-attribute half-noise (luminance):

$$S = N = 0.5 \cdot \begin{bmatrix} 1 \\ G_{Col} \\ G_{Pha} \end{bmatrix}^T \cdot \left(\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \right) \quad (3-23)$$

- Cross-attribute full-signal (luminance–color):

$$S = 0.5 \cdot \begin{bmatrix} 1 \\ G_{Col} \\ G_{Pha} \end{bmatrix}^T \cdot \left(\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right) \quad (3-24)$$

$$N = 0.5 \cdot \begin{bmatrix} 1 \\ G_{Col} \\ G_{Pha} \end{bmatrix}^T \cdot \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \right)$$

- Cross-attribute half-noise (luminance–spatial phase, luminance as signal):

$$S = 0.5 \cdot \begin{bmatrix} 1 \\ G_{Col} \\ G_{Pha} \end{bmatrix}^T \cdot \left(\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \right) \quad (3-25)$$

$$N = 0.5 \cdot \begin{bmatrix} 1 \\ G_{Col} \\ G_{Pha} \end{bmatrix}^T \cdot \left(\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \right)$$

Color and spatial phase signals were treated as binary (0 or 1), reflecting their presence or absence. In contrast, luminance input was passed through a nonlinear compressive response function to approximate the saturation characteristics of early visual encoding (Hamilton et al., 2007):

$$I_{Lum}(x) = \frac{x^\gamma}{x^\gamma + \epsilon^\gamma} \quad (3-26)$$

where x is the normalized luminance intensity (ranging from 0 to 1), and γ and ϵ are free parameters controlling the nonlinearity and semi-saturation point, respectively.

The full parameter set is:

$$\Theta = \{A, \sigma_{inner}, G_{Col}, G_{Pha}, \gamma, \epsilon\}$$

The model was fit to the averaged behavioral data across all participants using maximum likelihood estimation. A single parameter set was used to jointly fit all experimental conditions, including both Experiment 2-1 and Experiment 2-2.

4.4.2. Result

To evaluate whether the proposed single-channel model can account for the observed segmentation behavior, we fitted the model to the averaged data across all

participants from both Experiment 2-1 and Experiment 2-2. The model, with a single set of parameters Θ , was applied uniformly across all task conditions without condition-specific adjustments. This approach allows us to assess whether a unified computation—based on gain-weighted integration of attribute signals and shared temporal correlation—can capture both the overall segmentation performance and the asymmetric interference patterns across attribute combinations.

The fitted curves are shown in Figure 3-14. Overall, the model achieved an Akaike Information Criterion (AIC) of 2614.56, indicating good parsimony given the complexity and breadth of the experimental conditions. In terms of descriptive accuracy, the model accounted for 87.5% of the variance in the observed data ($R^2 = 0.88$), reflecting a strong overall fit across both Experiment 2-1 and Experiment 2-2. The estimated parameter set was $\Theta = \{A, \sigma_{inner}, G_{Col}, G_{Pha}, \gamma, \epsilon\} = \{3.3, 0.15, 0.33, 0.25, 2.00, 0.32\}$. Beyond the overall fit, we also evaluated model performance separately for within- and cross-attribute conditions. The model performed particularly well in the within-attribute conditions ($R^2 = 0.92$), while performance in the cross-attribute conditions was slightly lower ($R^2 = 0.86$). These results suggest that the proposed gain-weighted mechanism captures the dominant structure of segmentation behavior across varying sources of interference. However, the actual integration mechanism in the cross-attribute condition may deviate from the strictly linear integration assumed in the model, pointing to the possibility of more complex or nonlinear interactions between feature channels. This modeling outcome provides a quantitative foundation for evaluating the plausibility and limitations of the single-channel integration framework, which we further consider in the following discussion.

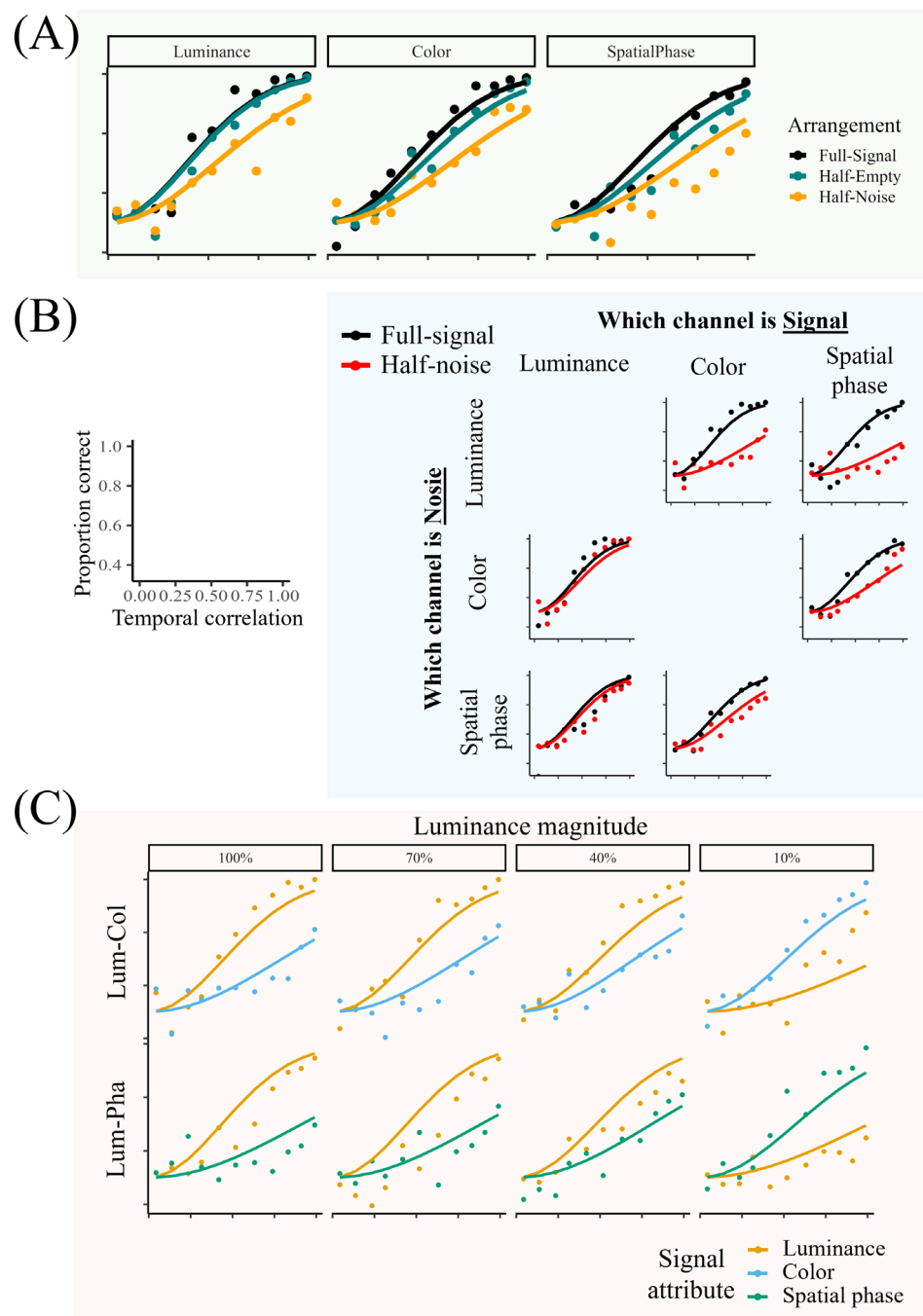


Figure 3-14. Model predictions and observed data across all experimental conditions. The proposed single-channel model, using a single parameter set Θ , was fit to the averaged behavioral responses from both Experiment 2-1 and Experiment 2-2. Dots represent empirical data, and solid lines represent model predictions. (A) Within-attribute conditions for luminance, color, and spatial phase. Each panel shows segmentation performance as a function of temporal correlation under three arrangements: full-signal (black), half-empty (teal), and half-noise (orange). This panel corresponds to Figure 3-9. (B) Cross-attribute interference conditions from Experiment 2-1. Each small plot shows the full-signal (black) and half-noise (red) conditions under different pairings of attributes. This panel corresponds to Figure 3-10. (C) Cross-attribute half-noise conditions from Experiment 2-2, where luminance magnitude was manipulated across four levels (100%, 70%, 40%, 10%). Each subplot shows model predictions and observed data for different signal attributes (orange: luminance, blue: color, green: spatial phase) under luminance-color (top row) and luminance-phase (bottom row) pairings. This panel corresponds to Figure 3-12.

4.5. General Discussion

The current study aimed to investigate how the human visual system integrates temporally correlated signals across different visual attributes during perceptual segmentation. While previous research has demonstrated that temporal correlation facilitates grouping within individual features, it remains unclear whether such computations are carried out independently within feature-specific channels or reflect a single unified channel across attributes. To address this question, we systematically manipulated temporal correlations across luminance, color, and spatial phase, and measured segmentation performance under different pairings of signal and noise attributes. The results show that noise from a different attribute can significantly impair segmentation, with an effect size comparable to that of noise from the same attribute. Furthermore, the magnitude of interference was not uniform across attributes: mutual interference levels varied depending on the relative internal signal gain between attributes. To account for these observations, we proposed a computational model based on the single-channel hypothesis. This model, which integrates signals across attributes using gain-weighted summation prior to temporal correlation, provided a sufficiently accurate fit to the behavioral data and captured the dominant structure of performance across all experimental conditions. In doing so, it offers a quantitative framework for evaluating the viability of a unified, gain-sensitive integration mechanism in temporal segmentation.

4.5.1. A unified gain-weighted mechanism accounts for cross-attribute temporal segmentation

This study first replicated prior findings showing that higher temporal correlation within a region enhances segmentation performance (Chen et al., 2025; S.-H. Lee & Blake, 1999). Extending this, we demonstrated that the visual system can simultaneously utilize multiple types of visual attributes—luminance, color, and spatial phase—for temporal segmentation, without a substantial decrement in performance (Chen et al., 2024a; Guttman et al., 2005). Additionally, consistent with previous work (Chen et al., 2025), we observed that segmentation based on luminance and color generally outperformed that based on spatial phase.

The novel contribution of the present study lies in two key observations. First, temporal correlations from different visual attributes not only combine to support segmentation but can also interfere with one another when noise is introduced in a separate feature channel. Second, the degree of such cross-attribute interference is not uniform: it depends on the relative internal strength of each attribute. Luminance, in particular, exerts stronger interference on color and spatial phase than vice versa. This asymmetry cannot be attributed to the physical stimulus properties, as all temporal streams were constructed with quasi-white noise spectra, ensuring that no attribute was inherently advantaged in terms of input signal structure. The observed asymmetry therefore reflects an internal property of the visual system, which we interpret as an attribute-specific gain mechanism.

Importantly, once feature-specific gain differences are taken into account, the asymmetry in internal signal strength across attributes can be effectively compensated. This is further supported by the results of Experiment 2-2, where gradually reducing the magnitude of luminance led to a systematic decrease in its interference effect on other attributes. These findings suggest that the relative strength of interference is not fixed, but depends on the internal representation strength of each attribute. Under this account, a single-channel model becomes sufficient to explain all major behavioral patterns observed across experiments. The model accurately captures both the facilitative effects of temporal correlation and the asymmetric interference between attributes, using a unified correlation computation applied to a gain-weighted integration of inputs. This supports the idea that the visual system may rely on a shared temporal integration mechanism, modulated by internal signal strength rather than hardwired feature segregation.

While we cannot completely rule out the possibility of more complex multi-channel architectures, our results suggest that such models may not be necessary to account for the present data. The gain-weighted single-channel model provides a parsimonious and quantitatively robust explanation, capturing both overall segmentation performance and asymmetric interference effects without requiring channel-specific computations.

4.5.2. The integration mechanism needs further research

The single-channel hypothesis assumes that all attribute signals—whether originating from the same or different visual features—are integrated into an amodal representation for further temporal processing and/or decision making. The concept of amodal signal comparison has been suggested by prior studies (Chen et al., 2024a; Fujisaki & Nishida, 2010; Nishida & Johnston, 2002), primarily in the context of local signal estimation. Our results extend this framework by applying it to a more complex segmentation task that requires global spatial integration.

A further step we took was to evaluate whether a simple linear summation across feature channels can account for performance in an amodal segmentation context. The results suggest that while such a model can replicate the overall behavioral trends, systematic discrepancies remain—particularly in Experiment 2-2, where the magnitude of the luminance signal was manipulated. Although the source of this discrepancy is not directly identifiable, a comparison of cross-attribute conditions across Experiments 2-1 and 2-2 reveals that the model performs slightly better in Experiment 2-1. This implies that a fixed gain ratio between attributes may suffice under constant conditions (as in Experiment 2-1), but begins to fail when the relative signal strengths are dynamically varied (as in Experiment 2-2), particularly in different pairings such as luminance–color and luminance–spatial phase. This observation hints at the possibility that the integration of feature signals into an amodal representation may involve nonlinear mechanisms. Such nonlinear integration could follow a classical Filter–Rectify–Filter (FRF) structure, wherein early-stage feature-specific signals are first filtered, then nonlinearly transformed, and subsequently re-integrated for higher-order perceptual decisions (DiMattina & Baker, 2021; Kingdom et al., 2003; Meese, 2010). Future work will be needed to explore the specific form of these nonlinearities in temporal integration across attributes.

Another open question concerns the source of the gain modulation itself. In our model, gain functions as a computational construct that directly modulates the internal strength of feature signals, but its underlying mechanism remains unspecified. Ideally, such a mechanism would allow for selective enhancement—or attenuation—of signals based on feature identity. One plausible candidate is feature-based attentional modulation. Prior studies have shown that attention can independently enhance sensitivity to

luminance and color features, thereby improving discrimination thresholds in a channel-specific manner (Liu, 2019; Morrone et al., 2002). These findings support the view that attentional resources may be allocated in a feature-selective fashion, dynamically adjusting the effective gain of different channels depending on task relevance or stimulus salience.

4.5.3. The neuron implementation of amodal integration

While the present study provides behavioral and computational evidence supporting the idea that temporal segmentation is more likely to be processed via a highly integrated channel spanning multiple visual attributes, we do not yet identify the corresponding neural mechanisms. Several brain regions have been implicated in related functions: the inferior parietal cortex and insula have been associated with temporal relationship processing, particularly temporal synchrony and asynchrony (Battelli et al., 2003, 2007; Bushara et al., 2001), while the intraparietal sulcus has been proposed to support multi-attribute integration into amodal visual representations (Roelfsema, 2006). However, the precise correspondence between the processes engaged in our task and their neural substrates remains to be established.

One promising approach to address this issue is visual adaptation. By selectively adapting the visual system to temporal correlations with distinct statistical profiles—such as specific eigenmodes or combinations of feature channels—it may be possible to uncover the presence of amodal, feature-insensitive encoding mechanisms. For example, if adaptation effects generalize across attributes, or if cross-feature aftereffects are observed, this would provide behavioral evidence for a shared integrative code. Conversely, the absence of such transfer would suggest that integration occurs downstream, after feature-specific processing has been completed.

Future studies that combine adaptation paradigms, neuroimaging, and computational modeling may offer critical insight into whether the gain-weighted integration observed behaviorally is supported by a unified neural substrate or emerges from a distributed functional architecture.

Chapter 5. Feature-Invariant Segmentation Emerges from Global Temporal Correlation

Combining the findings from Chapters 3 and 4, we established that a feature-invariant mechanism likely operates through a single information processing channel to support multi-attribute temporal segmentation. In this fifth chapter, we shift the focus toward the computational perspective of this ability. Specifically, we aim to examine how the visual system computes temporal correlation and how these temporal signals are transformed into perceptual grouping and segmentation.

5.1. Introduction

Visual perceptual organization is a fundamental function of the human visual system that enables the grouping of objects through various attributes, such as luminance, color, and temporal dynamics. These attributes were captured locally by different neurons, given that each neuron has a limited receptive field. To segment out a coherent object from others, an additional process is required to compare the local neuronal signals across space to determine whether they belong to the same object. One of the computational principles underlying this process is similarity-based grouping, a concept introduced in Gestalt psychology, which suggests that elements with similar features tend to be grouped together (Köhler, 1930; Wertheimer, 1938). Similarity operates across a wide range of visual attributes, both spatial and temporal. Spatial cues include orientation, color hue (Callaghan et al., 1986; Dakin & Watt, 1997; Michael S. Landy & Graham, 2003), luminance (Dakin, 2015), and motion properties such as direction and speed (i.e., common fate) (Sekuler, 1990; Uttal et al., 2000), while temporal cues involve temporal synchrony (Kojima, 1998; Rideaux et al., 2016; Usher & Donnelly, 1998), temporal structure (Blake & Lee, 2000; Chen et al., 2024a; S.-H. Lee & Blake, 1999), and generalized common fate (Sekuler & Bennett, 2001). The current study examined the computational mechanism for similarity-based grouping based on the temporal correlation of local feature change, although our argument may be generalized to similarity-based grouping based on other cues.

Computationally, whether a pair of two local elements are to be grouped or not depends not only on the similarity of that pair but also on the similarities of the other

element pairs. For example, consider the situation where four elements (A, B, C, and D) are spatially arranged as in Figure 5-1, and the pairwise similarity is high within {A, B} and within {C, D}, while relatively low for the other element pairs. The computationally reasonable solution in this case is to segregate the four elements into two groups {A, B} and {C, D}. To draw this solution, the visual system has to first evaluate the pairwise similarities between all local elements and then derive a global solution that maximizes within-group similarities while minimizing cross-group similarities. The computation based on the global similarity structure is simple when the number of elements is small, as in Figure 5-1-(A). It however could be complex when the number of elements is large, since the number of element pairs grows exponentially with the number of elements, as in Figure 5-1-(B).

Various global computations have been proposed in computer vision (CV) to address visual segmentation. For instance, self-attention transformers, which leverage pairwise similarity across all elements, have demonstrated performance comparable to human grouping abilities (Mehriani & Tsotsos, 2023). Similarly, Conditional Random Fields (CRFs), a specialized form of Markov Random Fields (MRFs), have been successfully applied to model dependencies between local similarities, enabling high-quality image segmentation (Krähenbühl & Koltun, 2011). Additionally, graph-based methods, such as Eigendecomposition on the graph Laplacian (a linear transformation of the similarity matrix), have achieved state-of-the-art performance (Jianbo Shi & Malik, 2000; Sun et al., 2021). These methods also exhibit similarities to human perceptual strategies, particularly in motion-based segmentation tasks (Sun et al., 2023a, 2023b, 2025).

However, past studies on human vision did not consider how global similarity structure of the stimulus the grouping performance. They only manipulated either within-group similarity or cross-group similarity, not both at the same time. For example, Lee & Blake (1999), Morgan & Castet (2002) modulate various levels of temporal correlation of motion direction change within a target area (i.e., within-group similarities were modulated), with no control over the cross-group similarities between target and background (S.-H. Lee & Blake, 1999; Morgan & Castet, 2002). Other studies also modulate the within-group similarities while maintaining the cross-group similarities close to zero (Alais et al., 1998; Sekuler & Bennett, 2001; Usher & Donnelly, 1998).

One reason for the lack of past studies on the effects of global similarity structure on human perceptual grouping may be that independently changing different aspects of similarity, such as within- and cross-grouping similarities, is technically challenging since the change of a single element affects all the similarity between this and other elements. To overcome this problem, we developed a novel stimulus generation method that leverages a Vision Transformer (ViT) (Dosovitskiy et al., 2021), which enabled us to precisely manipulate pairwise similarities between all the pairs of elements. In our stimulus design, we focused on the temporal effect in visual segmentation and defined the similarity between pairs of elements based on the correlation of their temporal sequences. Benefiting from well-defined loss functions, our stimuli were free from spatial artifacts, allowing us to isolate and examine human performance in purely similarity-based visual segmentation.

This study has three main contributions. First, we proposed and evaluated a ViT-based stimulus generator with a customized loss function to precisely control pairwise element similarity. Second, using the generated stimuli, we conducted a human psychophysical experiment to examine how perceptual segregation is affected by within- and cross-group similarities. Finally, we introduced a graph cut model (partially based on previous work (Jianbo Shi & Malik, 2000)) as a simple segmentation framework to test whether computations based on the global temporal similarity structure can approximate human perceptual performance.

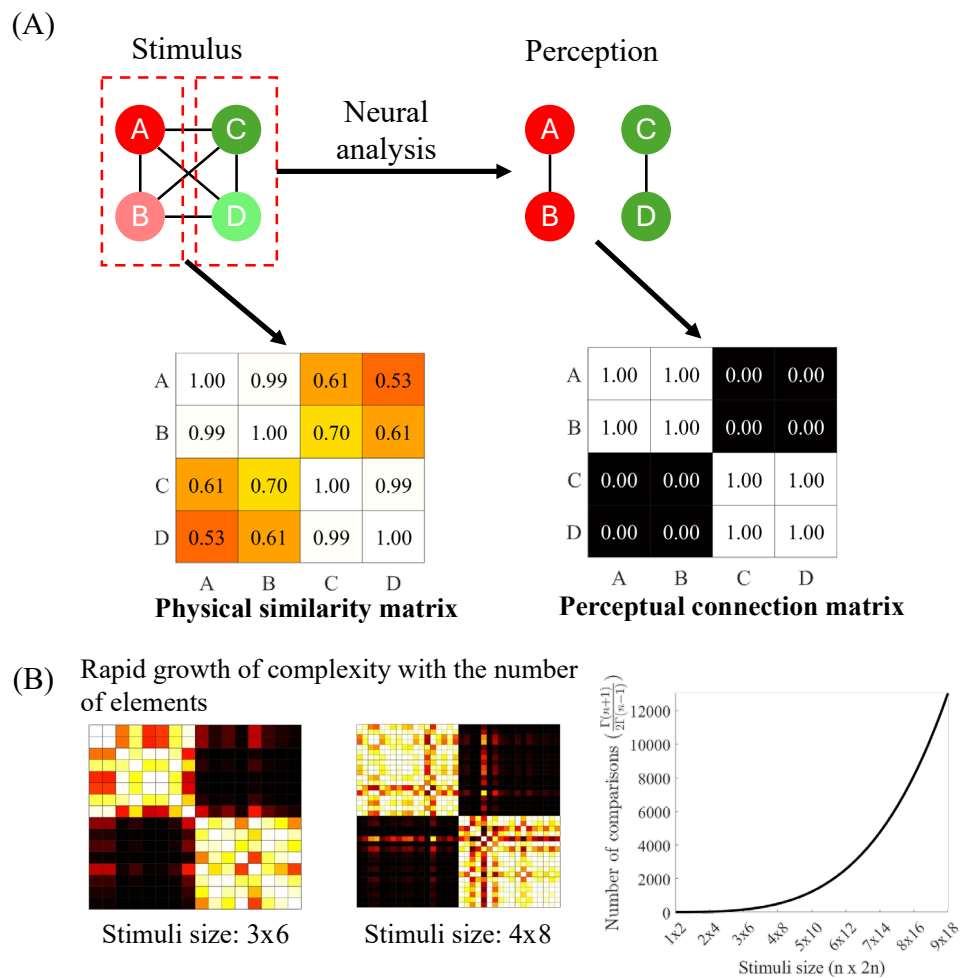


Figure 5-1. Linkage between similarity matrix and segmentation.

(A) Grouping/segmentation based on global similarity structure when the number of elements is small. A, B, C, and D are stimulus elements at different spatial locations. Assume that the physical pairwise similarities are high between A and B, and between C and D, while low for the other four pairs as shown by the physical similarity matrix. Potentially, this can yield the grouping result of {A,B} and {C,D} as shown by the dashed-red squares. After neural processing to consider the global similarity structure, similar pairs are grouped while dissimilar pairs are segregated as shown by perceptual connection matrix. (B) Demonstration of how the computational complexity increases with stimulus size. The left panel shows the physical similarity matrix for 3x6 and 4x8 element stimuli, respectively. The right panel shows how the number of pairwise comparisons rapidly increases with the number of elements.

5.2. Experiment 3-1. Stimuli Generation Incorporating Vit-Based Generator

5.2.1. Generation protocol

5.2.1.1. Feature tensor generated by visual transformer

In this section, we describe our approach to generating a 2D spatiotemporal stimulus using a lightweight vision Transformer. The goal of the model is to generate a feature tensor $F^* \in \mathbb{R}^{H \times W \times N}$ from a sampled noise tensor $F \in \mathbb{R}^{H \times W \times N}$ such that satisfies user-assigned cosine similarity (correlation) constraints both within (ξ_A or ξ_B) and across two areas (ξ_{cross}) (see Figure 5-2).

5.2.1.2. Problem definition

- We use a 2D feature map to represent spatiotemporal data. Consider a feature tensor F defined on a spatial grid G , let $H, W \in \mathbb{N}^+$ be the height and weight of the spatial grid. The spatial grid is defined $G = \{(i, j) | 1 \leq i \leq H, 1 \leq j \leq W\}$. At each spatial location (i, j) , there exists a corresponding temporal sequence $P_{ij} \in \mathbb{R}^N$. Here, N represents the length of the temporal sequence, which is set to 30 frames. F can be formally defined as follows:

$$F: G \rightarrow \mathbb{R}^N, \quad F(i, j) = P_{ij} \in \mathbb{R}^N, \quad \forall (i, j) \in G \quad (5-1)$$

- Splitting into sub-area. To impose specific correlation constraints, we partition the G into two subsets ϕ_A and ϕ_B :

$$\begin{aligned} \phi_A &= \left\{ (i, j) \mid 1 \leq i \leq H, 1 \leq j \leq \frac{W}{2} \right\}, \phi_B = \\ &= \left\{ (i, j) \mid 1 \leq i \leq H, \frac{W}{2} < j \leq W \right\} \end{aligned} \quad (5-2)$$

Here, we set H as 8, and W as 16. Therefore, the feature tensor is an 8×16 rectangle, with two 8×8 square sub-areas at the left and right sides.

5.2.1.3. Correlation constraints

We aim to generate F^* whose spatiotemporal patches exhibit assigned cosine similarities. Specifically, the similarities is calculated as following formula:

$$\text{sim}(\mathbf{P}_{ij}, \mathbf{P}_{i'j'}) = \frac{\mathbf{P}_{ij}^\top \mathbf{P}_{i'j'}}{\|\mathbf{P}_{ij}\|_2 \|\mathbf{P}_{i'j'}\|_2} \quad (5-3)$$

The similarity is calculated between the two vectors at spatial locations (i, j) and (i', j') . We want to realise the following properties:

- Within-area correlation. For any pair $((i, j), (i', j'))$ where both (i, j) and (i', j') belongs to the same area (ϕ_A , or ϕ_B), the similarity should be close to a desired constant ξ_A , or ξ_B , respectively. Formally:

$$\begin{aligned} \text{sim}(\mathbf{P}_{ij}, \mathbf{P}_{i'j'}) &\approx \xi_A, & \forall (i, j), (i', j') \in \phi_A, \\ \text{sim}(\mathbf{P}_{ij}, \mathbf{P}_{i'j'}) &\approx \xi_B, & \forall (i, j), (i', j') \in \phi_B \end{aligned} \quad (5-4)$$

Throughout this study, we kept ξ_A and ξ_B the same.

- Cross-area correlation. For any pair $((i, j), (i', j'))$ that spans different sub-areas, i.e., $(i, j) \in \phi_A$ and $(i', j') \in \phi_B$ (or vice versa), the similarity should be close to a distinct constant ξ_{cross} :

$$\text{sim}(\mathbf{P}_{ij}, \mathbf{P}_{i'j'}) \approx \xi_{cross}, \quad \forall (i, j) \in \phi_A, (i', j') \in \phi_B \quad (5-5)$$

Finding an analytical solution for F^* that satisfies all these constraints could be hard. Instead, we leverage a neural generative approach using a lightweight vision Transformer model.

5.2.1.4. Transformer-based generative model

Our generative model was based on a ViT backbone, incorporating convolutional residual blocks at both the encoding and decoding stages. The architecture can be summarized as follows:

- Encoder: A series of 2D CNN residual blocks (He et al., 2015) processed an initial latent tensor F , expanding the feature dimension from N to 128. In our implementation, we used several four stacked *ResidualBlock2D*

modules, each consisting of convolution, normalization (InstanceNorm2D), and *GELU* activation, with the residual connection between two layers.

- **Transformer layers:** We flattened the spatial dimensions and treated each location as a “token” in standard transformer terminology. Each token is represented by 128 dimension features, attended to every other token via multi-head self-attention (a similarity representation based on dot-product distance). Stacking multiple such layers enabled the model to capture global correlations across the entire feature map.
- **Decoder:** Another three layers of convolutional layers decode the features into the target spatiotemporal volume F^* , condense the feature dim from 128 to N . We applied non-linear activations ($\tanh$) to ensure an output within a controlled numeric range $(-1,1)$.

Mathematically, we defined the transformer layers as a function $\mathcal{T}(\cdot)$. The overall generation process transform a white-noise prior into target stimuli, could be expressed as:

$$F^* = \text{Dec}\left(\mathcal{T}(\text{Enc}(F))\right) \quad (5-6)$$

The model parameters were optimized so that the resulting output F^* fulfilled the cosine similarity constraints defined previously. The transformer was selected because the self-attention mechanism in transformers computes pairwise correlations (via the Query, Key, and Value operations) between all tokens in a layer. This is precisely the structural property we intended to exploit, aiming to modulate correlations (cosine similarity) within and across spatial locations in ϕ_A and ϕ_B . By stacking multiple transformer layers, the model could iteratively adjust and refine these pairwise relationships, converging on a solution that respects the target within-area and cross-area similarities.

5.2.1.5. Optimization

We optimized the model parameters to produce a stimulus F^* , which enforced both the within-area (within ϕ_A and ϕ_B) and cross-area correlations described earlier while also constraining the signal to have similar mean and variance values across areas.

The input for the generative model was initialized as a white noise tensor, i.e., random values drawn from a normal distribution with zero mean and moderate variance, encouraging broad coverage of frequencies.

To optimize the stimulus F^* , we defined a customized loss function $\mathcal{L}_{total}$ composed of several terms, each constrain a different statistical property:

- **Within-area correlation constraints:** We computed the cosine similarity among all pairs of vectors within each sub-area. This resulted in two similarity matrices, $S^A, S^B \in \mathbb{R}^{(H \times \frac{W}{2}) \times (H \times \frac{W}{2})}$, where S^A and S^B represent similarity values within the regions ϕ_A and ϕ_B , respectively. The within-area loss is defined as the summation of the absolute differences between the computed similarities and the desired similarity constants ξ_A and ξ_B . The within-area correlation loss $\mathcal{L}_A$ and $\mathcal{L}_B$ were then:

$$\mathcal{L}_A = |||\mathbf{S}^A| - \xi_A||_1, \quad \mathcal{L}_B = |||\mathbf{S}^B| - \xi_B||_1 \quad (5-7)$$

- **Cross-area correlation constraints:** The cosine similarity tensor $S^{AB} \in \mathbb{R}^{H \times \frac{W}{2} \times H \times \frac{W}{2}}$ was computed between the space spanning the areas ϕ_A and ϕ_B . The cross-area correlation loss $\mathcal{L}_{AB}$ was then:

$$\mathcal{L}_{AB} = |||\mathbf{S}^{AB}| - \xi_{cross}||_1. \quad (5-8)$$

- **Mean and variance regularization:** To maintain the whole tensor following a similar feature distribution, we added a loss to let the generated stimuli converge on specific statistics. Let $F^A, F^B \subset F$ denote the subsets of F corresponding to ϕ_A and ϕ_B , respectively. We aimed to match both global mean (μ_G) and variance (σ_G^2) values, defined as:

$$\begin{aligned} \mathcal{L}_{D,A} &= |\text{Var}(F^A) - \sigma_G^2| + |\text{Mean}(F^A) - \mu_G| \\ \mathcal{L}_{D,B} &= |\text{Var}(F^B) - \sigma_G^2| + |\text{Mean}(F^B) - \mu_G|. \end{aligned} \quad (5-9)$$

$\text{Var}(\cdot)$ and $\text{Mean}(\cdot)$ computed a single scalar for variance and the mean for the whole input tensor, respectively. Here, we set μ_G as 0 and σ_G^2 as $\frac{1}{3}$ to approximate a uniform distribution between -1 and 1 .

- **Equalizing global statistics:** To ensure that the feature distributions across ϕ_A and ϕ_B exhibit **similar mean and variance statistics**, we enforced

constraints on the mean and variance of each sub-region. We defined an additional penalty:

$$\mathcal{L}_{D,AB} = |\text{Mean}(F^A) - \text{Mean}(F^B)| + |\text{Var}(F^A) - \text{Var}(F^B)| \quad (5-10)$$

- **Variance expansion (“white noise” tendency):** To increase variability within each sub-area and push the solution towards a white noise-like distribution, we applied a negative penalty on intra-area pairwise distances:

$$\begin{aligned} \mathcal{L}_{V,A} &= -\text{Mean}\|F_{(i,j,:)}^A - F_{(i,j,:)}^A\|_1 \\ \mathcal{L}_{V,B} &= -\text{Mean}\|F_{(i,j,:)}^B - F_{(i,j,:)}^B\|_1 \end{aligned} \quad (5-11)$$

By negatively weighting the average pairwise differences, we encouraged the model to spread out feature values, thereby increasing variance.

- **Final loss function:** Combining all of the above components, we obtained the total loss as follows:

$$\begin{aligned} \mathcal{L}_{\text{total}} &= \alpha_1(\mathcal{L}_A + \mathcal{L}_B) + \alpha_2\mathcal{L}_{AB} + \alpha_3(\mathcal{L}_{D,A} + \mathcal{L}_{D,B}) + \alpha_4\mathcal{L}_{D,AB} \\ &\quad + \alpha_5(\mathcal{L}_{V,A} + \mathcal{L}_{V,B}) \end{aligned} \quad (5-12)$$

where α_k are scalar weights to balance the contributions of each term.

5.2.1.6. Training procedure

The network parameters (including those of the transformer and the encoder/decoder blocks) were optimized jointly via an AdamW optimizer. At each iteration, we:

- Generated an output stimulus F^* from the current model.
- Computed all required similarity, variance, and mean statistics to obtain $\mathcal{L}_{\text{total}}$.
- Performed backpropagation to update model parameters.

Since we began with white noise initialization and imposed correlation constraints alongside variance/mean regularizations, the final output remained statistically rich (i.e., maintained high variability) while respecting the assigned within- and cross-area correlation targets. This approach produced a refined F^* that approximated the ideal solution to our correlation objectives. We generated 30 different F^* values for each correlation pair.

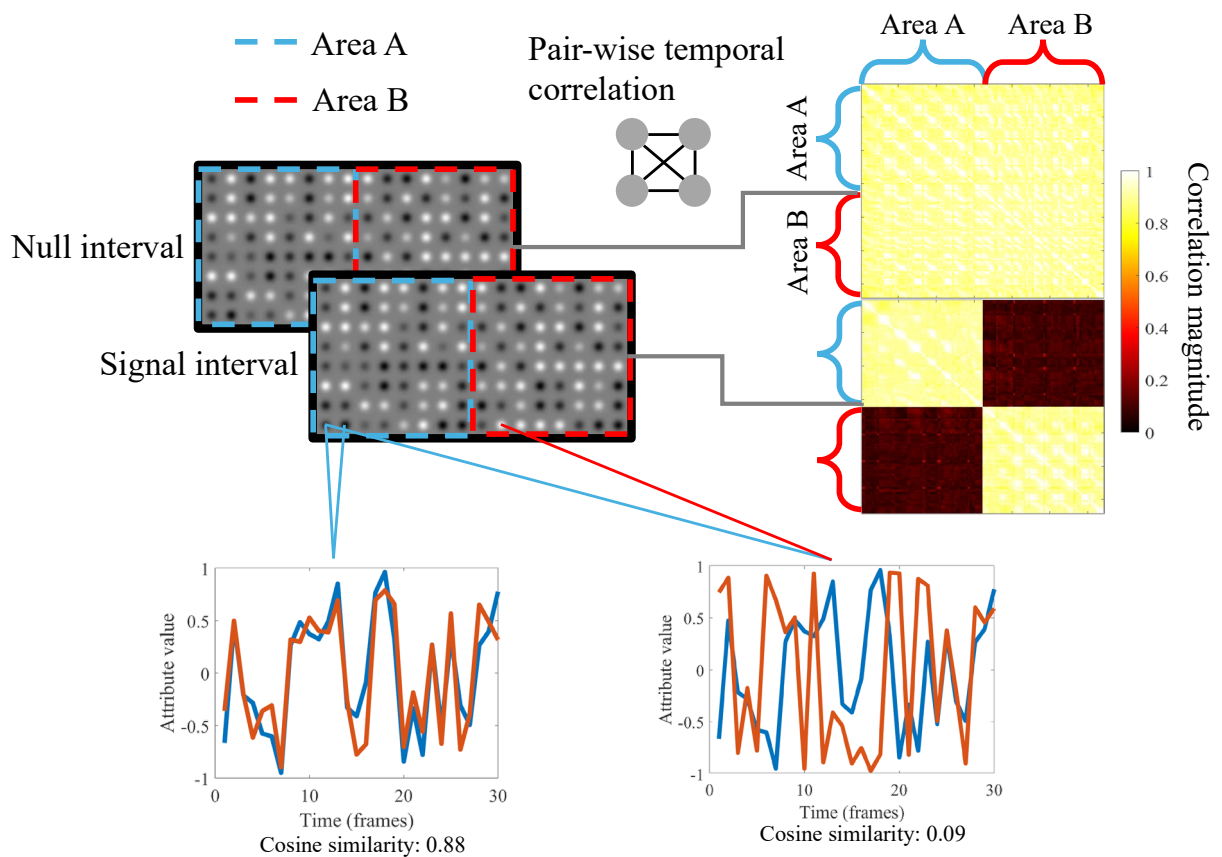


Figure 5-2. illustration of the experiment stimuli.

The luminance bulb is shown as an example here, but in the actual experiment, the stimulus could be a luminance bulb, color bulb, or luminance Gabor. The pairwise similarity heatmap is shown on the right; the space is flattened into a single dimension for convenient plotting. The position on the matrix corresponding to the 2D location is labeled, with the blue part corresponding to Area A and the red part corresponding to Area B. A demo time series is plotted at the bottom; the left figure demonstrates the elements with high temporal similarity, while the right figure demonstrates those with low similarity. The demo video can be viewed at https://osf.io/9yncd/?view_only=eb481f2b28ab42e29ff267833f70b8e5

5.2.2. Result

5.2.2.1. Invisibility of spatial cues

To confirm that the generated stimuli could only be distinguished based on the designated temporal similarity structure, we first verified that they were unlikely to be differentiated by other spatial indicators. We calculated two standard measures of spatial cues: the spatial average value per frame, representing the attribute's convergent value (e.g., mean luminance), and the spatial root mean square deviation per frame, indicating the attribute's concentration magnitude (e.g., luminance contrast). These

statistics were selected based on previous studies demonstrating that the human visual system is particularly sensitive to such regularities when evaluating texture (Dakin, 2015; Sikder et al., 2024).

Analysis of the first index revealed that the mean luminance difference between the left and right partitions (Areas A and B in Figure 5-2) across 30 frames was 0.17% for all correlation pairs, with a peak difference of 0.24%. This negligible discrepancy suggested that participants were unlikely to perceive a difference based on mean luminance alone. Similarly, for the second index, the findings indicated that the average root mean square deviation difference between partitions across 30 frames was 0.34% for all correlation pairs, reaching a maximum of 0.61%. This magnitude was well below the threshold required for human perception of such differences, confirming that spatial cues were effectively minimized.

5.2.2.2. Validation of cosine similarity distribution

Next, we assessed whether the stimuli generated by the ViT exhibited the designated cosine similarity values. Figure 5-3 presents the correlation distribution, obtained by directly computing the absolute cosine similarity of the generated stimuli. The figure demonstrates that both within-area and cross-area correlations closely align with the assigned correlation values. Furthermore, additional quantitative analysis revealed high predictability of the measured correlation based on the assigned correlation ($R^2 = 0.99$).

These results demonstrated that our generation protocol effectively manipulated the similarity structure of the stimuli as intended.

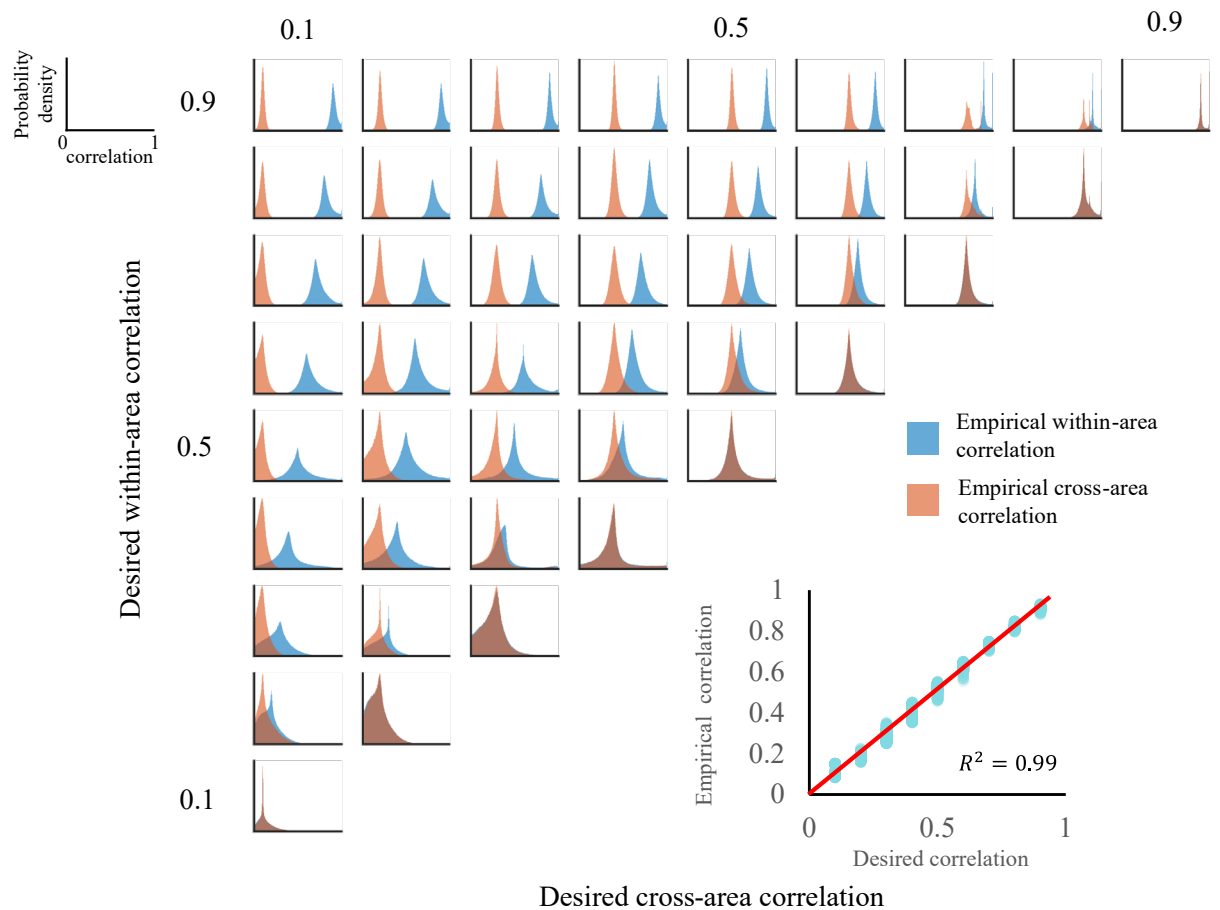


Figure 5-3. Distribution of temporal correlations in the generated stimuli.

The assigned within- and cross-area correlation values are shown on the axis labels. Each subgraph displays the empirical correlation distribution separately for within- and cross-area correlations, with within-area correlations represented in blue and cross-area correlations in orange. The lower-right quadrant presents a scatter plot of the desired correlation values (x-axis) versus the actual values (y-axis). The red line represents the identity line ($y = x$) as a reference.

5.2.2.3. Quasi-white noise characteristics

Finally, we analyzed the spectral properties of the generated stimuli. The temporal spectrum for Areas A and B was computed separately (Figure 5-4). Regardless of the correlation pair, the spectrum exhibited a near-uniform distribution, indicating that the generated stimuli closely resemble white noise. This broadband energy distribution ensures activation of diverse neural mechanisms. Additionally, there were no significant spectral differences between Areas A and B during the signal interval, suggesting that simple linear operations would be insufficient to distinguish between the two partitions.

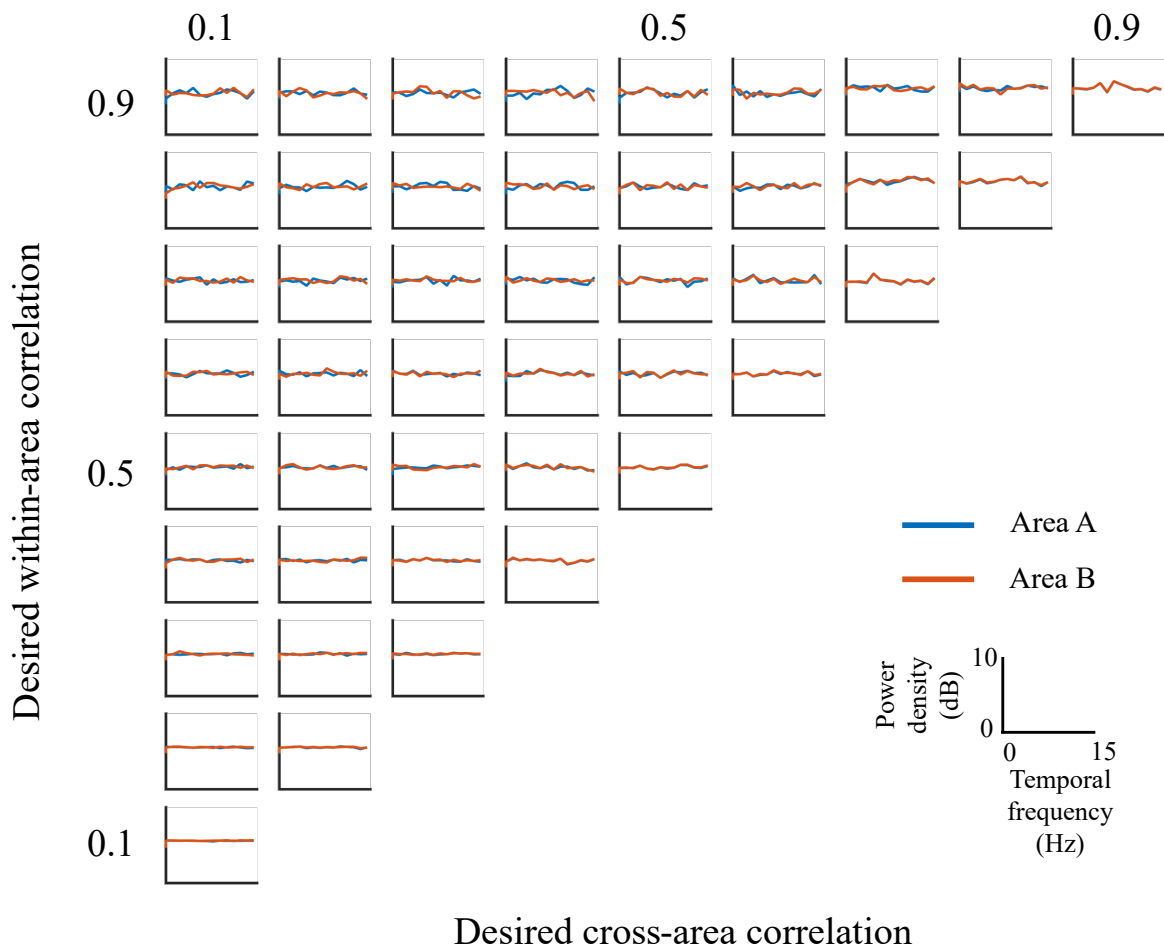


Figure 5-4. Temporal spectrum of the generated stimuli.

The assigned within- and cross-area correlation values are indicated on the axis labels. The mean power spectrum within Area A and Area B is shown in blue and orange, respectively. All subgraphs share the same axis scaling, as depicted in the lower-right section of the figure, where the x-axis represents temporal frequency (Hz) and the y-axis represents power density (dB).

5.2.2.4. Exclusion of higher-order visual feature visibility

Our stimulus design also excluded higher-order statistical features that could be used in the segmentation task, such as local motion differences and temporal pattern recognition.

Humans can perceive shape based on local motion differences (i.e., shape-from-motion) (Murray, 2003; Tangemann et al., 2024; Uttal et al., 2000). However, our stimulus design mitigates this possibility. Temporal similarity was computed using the absolute value of cosine similarity, meaning that simultaneous increases or decreases in brightness were treated as having the same magnitude of temporal similarity. This approach ensured that any potential motion signals were drift-balanced across the left and right partitions, eliminating systematic directional motion cues that could facilitate segmentation.

Additionally, humans can distinguish regions by recognizing different temporal regularities after long exposure or repeated presentations (Turk-Browne et al., 2005). However, our stimuli were designed to approximate quasi-white noise in the temporal domain, making it difficult to extract stable, discernible patterns in a single exposure. Given the short presentation duration and the absence of repeated trials with the same sequence, participants likely lacked sufficient information to develop reliable expectations about temporal structures.

5.3. Experiment 3-2. Human Segmentation Performance for The Generated Stimuli

In this section, we report the results of a psychophysical experiment designed to investigate the effect of within- and cross-group temporal correlations on human segregation performance.

5.3.1. Methodology

5.3.1.1. Apparatus

The stimuli were displayed on a VIEWPixx /3D LCD monitor (VPixx Technologies, Saint-Bruno-de-Montarville, Canada) with a resolution of $1,920 \times 1,080$ pixels and a refresh rate of 30 Hz. The lowest, mean, and highest display luminance ranged from 1.8 cd/m^2 (lowest) to 96.7

cd/m² (highest), with a mean luminance of 48.4 cd/m². The linear output of luminance was calibrated for each channel using an i1Pro chromometer (VPixx Technologies). Each pixel subtended 1.3 arcmin at a viewing distance of 70 cm. Participants were seated in a dark room with a chinrest to stabilize head position. The experiments were programmed using the Psychophysics Toolbox (Brainard, 1997) in MATLAB release 2022a (MathWorks, Natick, MA, USA). The study protocol adhered to the ethical standards of the Declaration of Helsinki, except for preregistration, and was approved by the Ethics Committee of Kyoto University (KUIS-EAR-2020-003).

5.3.1.1. Participants

The experiment included two of the authors and three naïve participants (four males; mean age: 24.67 years), all of whom reported normal or corrected-to-normal vision. Participants were informed of the study's nature and provided written informed consent before participation. Monetary compensation was provided upon completion. Prior to the main experiment, a revised version of the minimum motion method (Anstis & Cavanagh, 1983; Cavanagh et al., 1987) was used to measure the equiluminant ratio between the red (R) and green (G) channels for each participant. Participants viewed a rightward-drifting Gabor, where the R and G channels were 180° out-of-phase, and adjusted the channel ratio to minimize the perception of motion. The mean R:G ratio was determined to be 0.58.

5.3.1.2. Stimuli

The actual stimulus (demonstrated in Figure 5-2) was based on the generated sequence $F_{8 \times 16 \times 30}^*$, randomly drawn from the 30 variants. The stimulus sequence for one presentation was a dynamic 8×16 element lattice lasting for 30 frames (1 s). Each element subtended 0.5° in diameter, with the entire lattice subtending $4^\circ \times 8^\circ$. Each value in F^* was transformed into a $0.5^\circ \times 0.5^\circ$ element within the lattice, and all elements were concatenated, yielding the final stimulus. With luminance as the attribute, the element was defined as a Gaussian bulb as follows:

$$L(x, y, t) = L_{mean} \cdot \left(1 + F^*(i, j, t) \cdot e^{\frac{x^2 + y^2}{\sigma^2}} \right) \quad (5-13)$$

Where x and y are spatial locations on the plane, 0 is the center of the element, and t is the frame. i, j is the location within F^* , where i is an integer ranging from 1 to 8, and j

is an integer ranging from 1 to 16. L_{mean} is the mean luminance of the screen, and σ is the Gaussian windows of the element, set as 5.25 arcmin.

The element was also a Gaussian bulb with color as the attribute, but with three separate channel assignments:

$$L(x, y, t) = \begin{cases} R(x, y, t) = L_{mean} \cdot \left(1 + F^*(i, j, t) \cdot e^{\frac{x^2+y^2}{\sigma^2}}\right) \\ G(x, y, t) = L_{Iso.RG} \cdot L_{mean} \cdot \left(1 - F^*(i, j, t) \cdot e^{\frac{x^2+y^2}{\sigma^2}}\right) \\ B(x, y, t) = L_{mean} \end{cases} \quad (5-14)$$

where $L_{Iso.RG}$ is the iso-luminant index between red and green.

Taking spatial phase as the attribute, the element changed to the Gabor patch as follows:

$$L(x, y, t) = L_{mean} \cdot \left(1 + \cos\left(2\pi\omega r - \frac{\pi}{4}F^*(i, j, t)\right) \cdot e^{\frac{x^2+y^2}{\sigma^2}}\right) \quad (5-15)$$

$$r = y\cos(\theta) + x\sin(\theta)$$

where ω is the spatial frequency, set to 2 cpd, and θ is the orientation of the Gabor pattern, which was randomly drawn from the uniform distribution amongst $[0, 2\pi)$. The $\frac{\pi}{4}$ term in advance the F^* ensured that the spatial phase change could not exceed 90° per frame, thereby preventing any effect of luminance flickering.

5.3.1.3. Design and procedure

The experiment comprised multiple trials in which participants viewed two consecutive 1-second video clips: a signal interval and a null interval (Figure 2). The signal stimulus exhibited varying degrees of within- and cross-area temporal correlation, quantified by the absolute value of cosine similarity, whereas the null stimulus did not. For each trial, participants were asked to decide which of the intervals is able to separate into two square halves by pressing the keyboard (left arrow for first interval, right arrow for second interval). The proportion correct was calculated as the percentage of trials in which participants correctly identified the signal interval. Throughout the experiment, we tested nine different levels of cosine similarity, ranging from 0.1 to 0.9 in increments of 0.1, for both within- and cross-area correlations. To assess the generalizability of the

similarity structure in visual segmentation, we introduced three different attributes: luminance, color (red/green), and spatial phase.

In each trial, participants sequentially viewed two 1,000-ms video clips, separated by a 100-ms interval. One clip, designated as the signal interval, was distinguishable due to the difference between within- and cross-area correlations, while the other, referred to as the null interval, lacked this distinction. The within-area correlation remained constant across both clips, with only the strength of the correlation between within- and cross-area elements varying. The order of presentation for these clips was randomized across trials. Participants were instructed to identify the interval in which the texture appeared divided into two distinct parts. Using a keyboard, they indicated whether the first or second interval contained the signal stimulus. Each participant completed 30 segmentation trials for each pair of within- and cross-area correlations, with segmentation performance assessed based on the proportion of correctly identified signal intervals.

We tested 45 different correlation pairs in which the within-area correlation was either higher than or equal to the cross-area correlation. When the cross-area correlation exceeded the within-area correlation, segmentation was not possible, as the visual distinction between areas was lost.

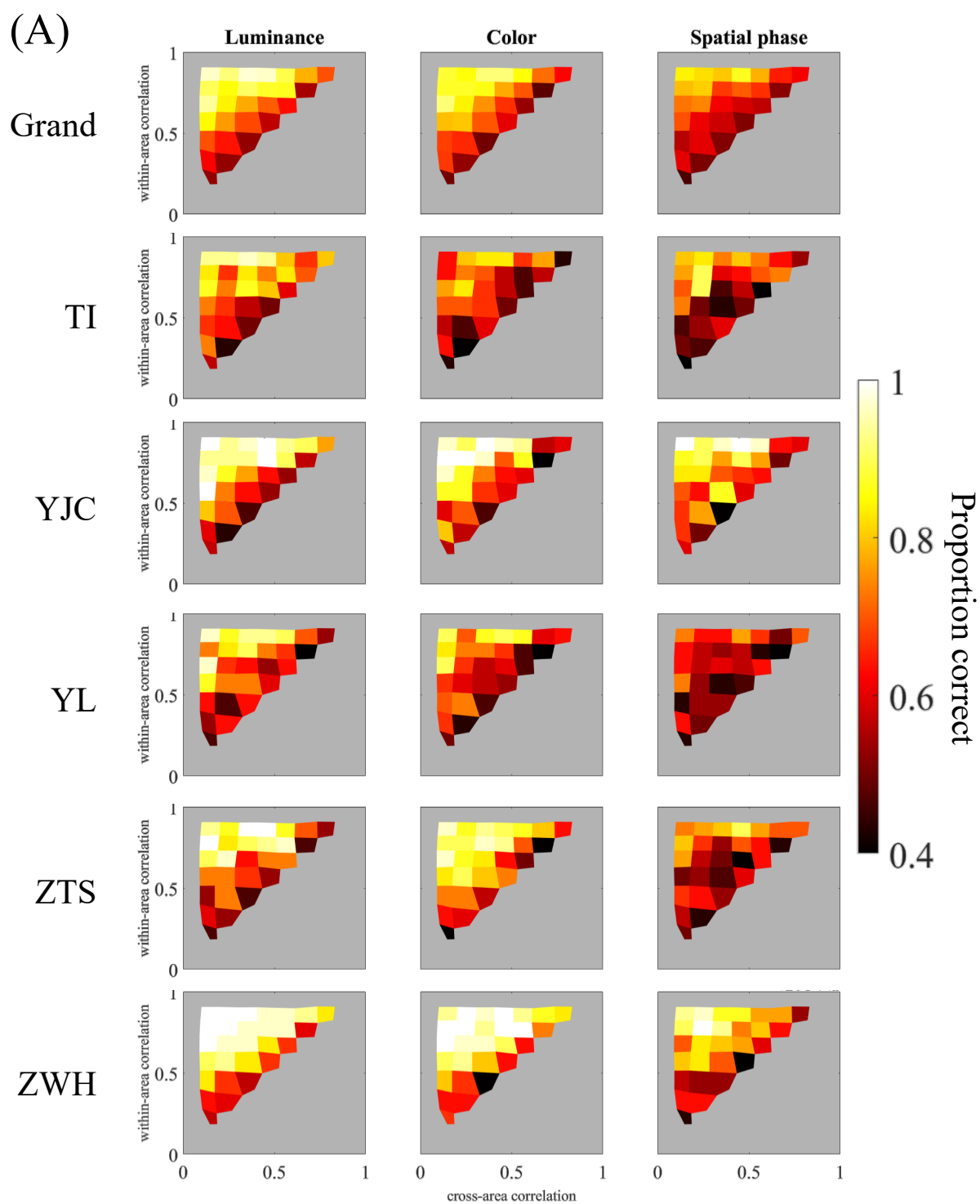
To ensure comprehensive evaluation, we examined three attributes: luminance, color, and spatial phase. The order of these attributes, as well as the sequence of correlation conditions within each attribute, was randomized. To complete all conditions, comprising three attributes and 45 correlation pairs, participants dedicated approximately 7.5 hours to the experiment.

5.3.2. Result

Figure 5-5-(A) illustrates the participants' segmentation accuracy, expressed as the proportion of correct responses, across different correlation pairs and for the three visual attributes: luminance, color, and spatial phase. We also show the contour plot version in Figure 5-5-(B) to show the general tendency. A consistent trend was observed across all attributes, with segmentation performance varying as a function of both cross-area and within-area correlations. Specifically, segmentation perception improved as the within-area correlation increased and the cross-area correlation decreased. We show

the contour graph here, while the original heatmap can be found in the supplementary material.

The results indicated that visual segmentation performance improves only when the difference between within- and cross-area correlations is sufficiently large. Although the general trend remained consistent across attributes, we observed variations in overall accuracy, with luminance yielding the highest accuracy, followed by color and then spatial phase. This suggested that segmentation based on temporal similarity structure is more challenging when the distinguishing attribute is spatial phase compared to luminance or color. Additionally, the contour of segmentation performance in Figure 5-5 deviated slightly from parallel alignment with the diagonal line (where the within-area correlation equals the cross-area correlation). This suggested that within-area and cross-area correlations do not symmetrically influence segmentation performance. A possible explanation for this phenomenon may lie in the spatial arrangement of the cross-area pairs, where the mean spatial distance is twice that of the within-area pairs.



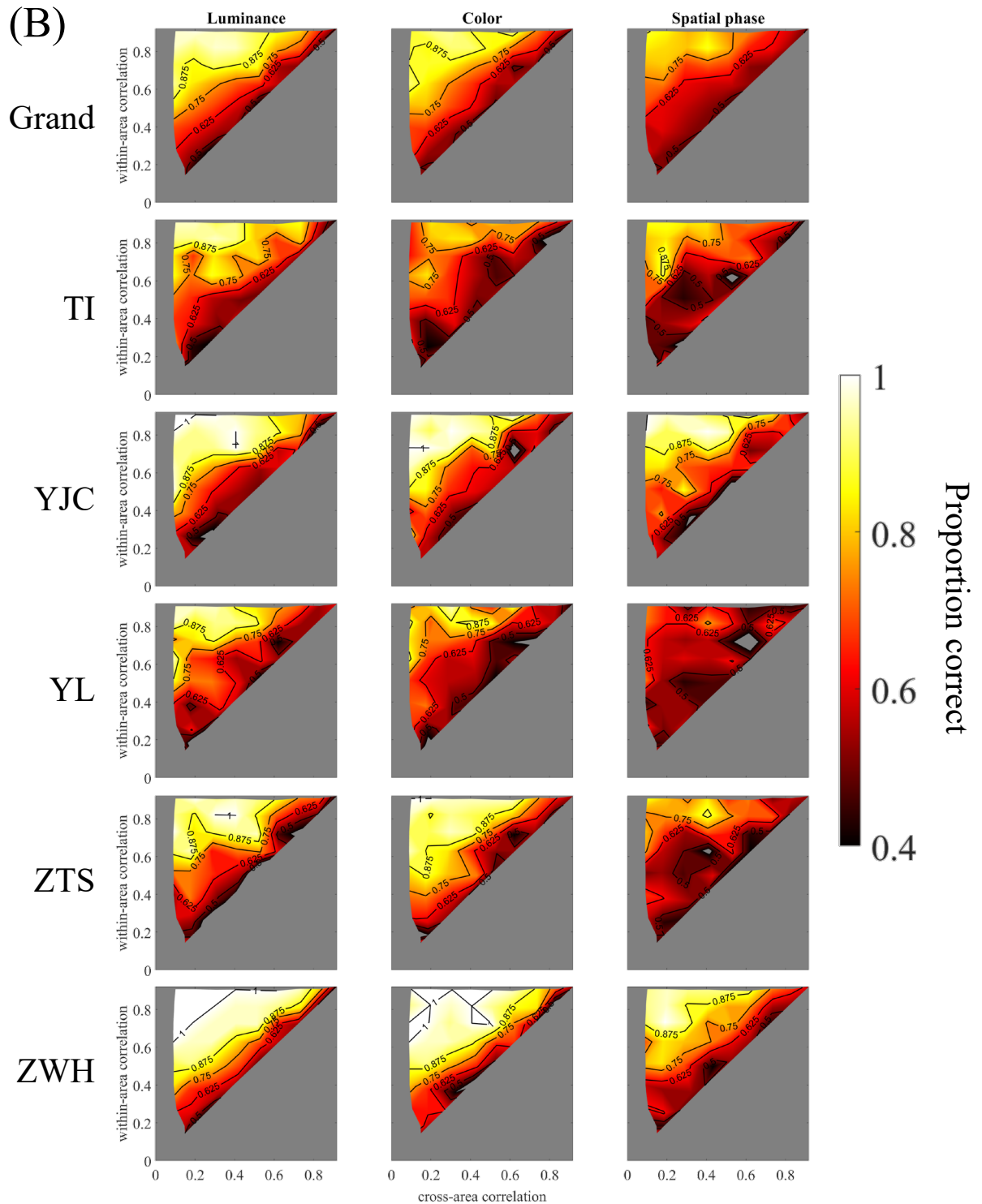


Figure 5-5. Human segmentaion performance restult.

(A) Perceptual segmentation results based on luminance, color, and spatial phase. The first row presents the grand average results across five participants, while individual participant data are shown in subsequent rows. Columns represent the results for the luminance, color, and spatial phase conditions. Within each graph, the x-axis represents the empirical cross-area correlation of the generated stimuli, while the y-axis represents the empirical within-area correlation. The proportion of correct responses is represented by reddish intensity, with contour lines included. The gray regions outside the triangular area indicate untested conditions, while conditions within the triangular area have values < 0.4 . (B) the contour plot version upon (A) to show the general tendency of the performance

5.4. Experiment 3-3. Analysis of Human Performance Using a Graph Cut Model

The results presented in the previous section have indicated that human visual processing incorporates both within- and cross-area correlations in similarity-based grouping, demonstrating its ability to process intricate similarity structure cues to infer spatial arrangements. However, the similarity structure alone does not explicitly define which group or object an individual pixel should be assigned to. Consequently, the visual system must employ an algorithm to convert similarity structure cues into predictions regarding object or group assignments without supervision, meaning it does not rely on predefined correct answers.

To address this challenge, we utilized a classical graph cut model to simulate segmentation performance for our stimuli and assessed the extent to which the model can replicate human performance. Figure 5-6 illustrates the simulation protocol and model architecture. Similar to human participants, the model processed two video clips, a signal interval, and a null interval, and determined which clip can be divided into left and right spatial partitions. The model underwent multiple trials, and segmentation accuracy was measured as the proportion of correctly identified signal intervals. The proportion correct was computed across all 45 correlation pairs tested in human experiments.

There were two model architecture versions: a naïve model and a full model. The naïve model did not include any trainable parameters and followed a three-stage processing sequence to evaluate the inherent capabilities of the existing structure. First, the model computed a pairwise temporal cosine similarity matrix, reflecting the experimental finding that human visual processing relies on similarity structures to infer spatial configurations. Next, it performed eigendecomposition on the similarity matrix to generate a spatial segmentation representation, thereby identifying group memberships for each pixel in an unsupervised manner. Finally, the model compared the estimated group memberships with the true ones and selected the interval (signal or null) maximizing similarity.

In the full model, we introduced two adjustable parameters (Figure 5-6) to refine the model's performance and better align it with human capabilities. The first parameter incorporated noise before computing the similarity matrix, with the signal-to-noise ratio

(SNR) controlling the noise level. Noise is an inherent aspect of human visual processing and disrupts similarity computations, leading to variability in perceived similarity strength. In our model, both noise addition and similarity computation were linear operations, meaning that the order of these steps—whether noise is introduced before or after similarity computation—did not affect the final output.

The second parameter accounted for the possibility that the visual system may not evaluate all similarity pairs equally. Previous studies (Maruya et al., 2013; Motoyoshi, 2004) have suggested that the ability to judge simultaneity declines as the spatial distance between elements increases, potentially due to longer signal transmission times or reduced neural connectivity between distant spatial locations. To replicate this effect, we introduced a distance-dependent weighting function in the similarity matrix, reducing the contribution of similarity strength for elements that are farther apart. The position-embedding sigma controlled this weighting as a function of distance. A higher sigma value allowed the model to consider more widely separated element pairs, whereas a lower sigma value restricted similarity computations primarily to neighboring elements, such as those at the segmentation boundary.

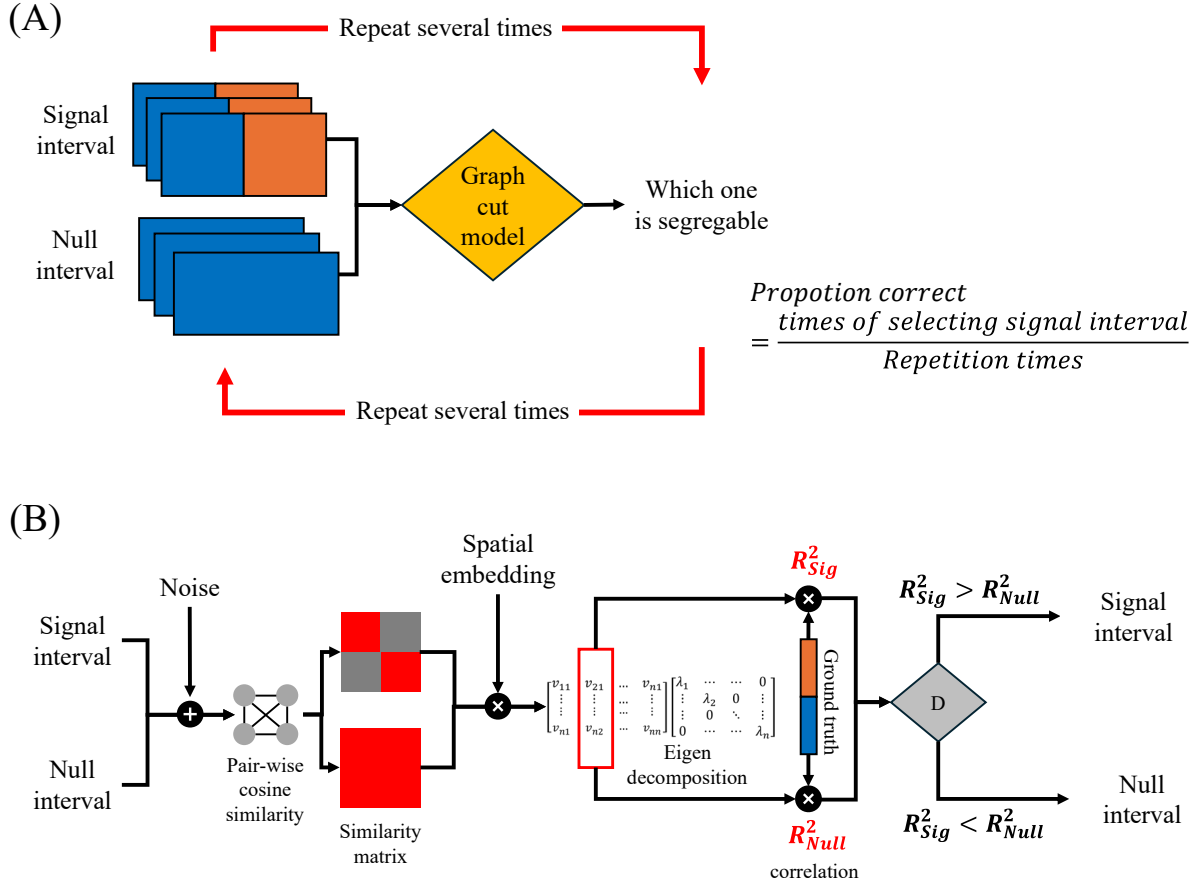


Figure 5-6. Model simulation protocol and architecture

(A) Simulation protocol for measuring model visual segmentation performance. (B) Graph cut model architecture.

5.4.1. Methodology

5.4.1.1. Model architecture

This section presents the model architecture tested in the simulation. The model utilized temporal correlation to construct a fully connected undirected graph across spatial locations, finding the coarsest subgraph as the estimation of segmentation.

The model input stimuli of the same size as those generated by the ViT ($8 \times 16 \times 30$). Two sequences, signal interval (F_{sig}^*) and null interval (F_{Null}^*), were simultaneously used as model input. In most cases, the calculations applied to both the signal and null intervals. Unless otherwise specified, only the computations for signal interval are presented, with the null interval being processed in the same manner.

After the stimuli were input, we added internal noise to the input as follows:

$$F'_{sig} = \frac{\tau}{1 + \tau} \cdot F_{sig}^* + \frac{1}{1 + \tau} \cdot \mathcal{E} \quad (5-16)$$

where $\mathcal{E}$ is Gaussian white noise with the shape $8 \times 16 \times 30$, a mean of 0, and variance of 1. τ represents the free parameters controlling the SNR within the range of $[0, \infty)$.

After adding noise, we flattened F'_{sig} from 3D to 2D, resulting in X_{sig} with dimensions $(8 \times 16) \times 30$. The mapping relationship was as follows:

$$F'_{sig}(k, t) = X_{sig} \left(\left\lceil \frac{k}{8} \right\rceil, \text{mod}(k, 8), t \right) \quad (5-17)$$

$$k \in \{1, 2, \dots, 96\}$$

where $\lceil \cdot \rceil$ is the ceiling function, returning the smallest integer greater than or equal to the argument, and where $\text{mod}(\cdot)$ returns the remainder after division.

Next, we calculated the similarity matrix based on X_{sig} . The cosine similarity between any pair of vectors i and j was given by:

$$S_{sig}(i, j) = \left| \frac{X_{sig}(i, :) X_{sig}(j, :)^T}{\|X_{sig}(i, :)\|_2 \|X_{sig}(j, :)\|_2} \right| \cdot \Omega(i, j), \quad (5-18)$$

$$\forall i, j \in \{1, 2, \dots, 96\}$$

$$\Omega(i, j) = e^{-\frac{\Delta(i, j)^2}{\sigma_{\Delta}^2}}$$

$$\Delta^2(i, j) = \left(\left\lceil \frac{i}{8} \right\rceil - \left\lceil \frac{j}{8} \right\rceil \right)^2 + (\text{mod}(i, 8) - \text{mod}(j, 8))^2$$

Here, the similarity was computed as the absolute value of the cosine similarity, weighted by a Gaussian function $\Omega(i, j)$, which reduced the connectivity between distant elements. The parameter σ_{Δ}^2 controlled the width of the Gaussian windows, with values in the range of $[0, \infty)$.

We then transformed the similarity matrix into an affinity matrix by applying a step function, turning it into a simple binary graph structure instead of a weighted graph:

$$A_{sig}(i, j) = \begin{cases} 1, & S_{sig}(i, j) \geq \overline{S_{sig}} \\ 0, & \text{otherwise} \end{cases} \quad (5-19)$$

where $\overline{S_{sig}}$ is the mean value of the whole similarity matrix.

Next, we performed a classical eigendecomposition on the normalized Laplacian matrix as follows:

$$\begin{aligned}
 L_{Sig} &= I - D_{Sig}^{-\frac{1}{2}} A_{Sig} D_{Sig}^{-\frac{1}{2}} \\
 D_{Sig} &= \text{diag} \left(\sum_{j=1}^{8 \times 16} A_{Sig}(i, j) \right) \\
 \text{diag}(V_{n \times 1}) &= V_{n \times 1} \odot I_{n \times n}
 \end{aligned} \tag{5-20}$$

The eigendecomposition of the Laplacian matrix yielded:

$$L_{Sig} = v_{sig}^{-1} \lambda_{sig} v_{sig} \tag{5-21}$$

where v_{sig} represents the eigenvectors and λ_{sig} represents the eigenvalues. According to a previous study (Fiedler, 1989), the eigenvectors corresponding to the second smallest eigenvalue (also called the Fiedler vector) is the optimal solution for partitioning the graph into two subgraphs. We transformed the vector into matrix form as follows:

$$G_{Sig}(h, w) = v_{sig}((h - 1) \times 8 + w, l), \tag{5-22}$$

where l is the position corresponding to the second smallest eigenvalue. This process provided the coarsest graph structure. In simple binary segmentation tasks, such as figure-ground segmentation and in our task, the value of the Fiedler vector is directly proportional to the probability that a given spatial point belongs to the target.

Finally, we transformed the eigenvector into a decision-making process based on the correctness rate. Specifically, the model outputs the option with the highest similarity to the answer. The correctness rate was defined as follows:

$$\begin{aligned}
 R_{Sig}^2 &= \left(\frac{\sum_h \sum_w (G_{Sig}(h, w) - \overline{G_{Sig}})(\mathbb{T}(h, w) - \overline{\mathbb{T}})}{\sqrt{\sum_h \sum_w (G_{Sig}(h, w) - \overline{G_{Sig}})^2} \sqrt{\sum_h \sum_w (\mathbb{T}(h, w) - \overline{\mathbb{T}})^2}} \right)^2 \\
 \mathbb{T}(h, w) &= \begin{cases} 1, & 1 \leq h \leq 8, 1 \leq w \leq 8 \\ -1, & \text{otherwise} \end{cases}
 \end{aligned} \tag{5-23}$$

where $\mathbb{T}(h, w)$ is the target matrix (i.e., the answer).

The model output the result based on the following decision rule:

$$D(R_{sig}^2, R_{Null}^2) = \begin{cases} signal, & R_{sig}^2 > R_{Null}^2 \\ null, & R_{sig}^2 < R_{Null}^2 \\ \gamma, & otherwise \end{cases} \quad (5-24)$$

$$\gamma \sim Bernoulli(p = 0.5)$$

where γ is a random variable following the Bernoulli distribution with equal probability of returning either signal or null.

5.4.1.2. Simulation protocol

For each trial, the model generated two possible outcomes: signal or null. This process was repeated 1,000 times, with input stimuli randomly selected from 30 variants of a single correlation pair for each iteration. Based on these 1,000 responses, we calculated the percentage of signal responses, defining it as the proportion correct. We tested all 45 correlation pairs to obtain the model's performance in the segmentation task, representing the proportion correct as a function of both within- and cross-area correlation (Ψ_{Model}). The model's goodness of fit was determined by its ability to predict human responses (Ψ_{Human}), calculated as follows:

$$R^2 = 1 - \frac{\sum_{\phi_{cross}} \sum_{\phi_{within}} (\Psi_{Model}(\phi_{cross}, \phi_{within}) - \Psi_{Human}(\phi_{cross}, \phi_{within}))^2}{\sum_{\phi_{cross}} \sum_{\phi_{within}} (\Psi_{Human} - \bar{\Psi}_{Human})^2} \quad (5-25)$$

where ϕ_{cross} , ϕ_{within} represent cross- and within-area correlation, respectively.

For the naïve model, the parameter was configured to have no impact on model performance, with the signal rate (i.e., $\frac{\tau}{1+\tau}$) set to 1, corresponding to an infinite SNR. Additionally, the position-embedding sigma (i.e., σ_{Δ}^2) was set to infinity, causing the weight function to become 1 universally. To determine the most suitable parameter pair for replicating the human result, we implemented a grid search sampling approach in the parameter space. We selected 50 samples with linear intervals from $0 \leq \frac{\tau}{1+\tau} \leq 1$ for the SNR. Similarly, for the position-embedding sigma, we chose 50 samples with linear steps from $0 \leq \sigma_{\Delta} \leq 16.5$, where the maximum value represents the largest Euclidean distance in our stimuli. This resulted in 2,500 parameter pairs being sampled across the parameter space, which were then used to assess the model's goodness of fit against human responses.

5.4.2. Result

5.4.2.1. Naïve model simulation results

Figure 5-7 presents the simulation results of the naïve model. The top section of the figure displays a single trial output from the second stage. Without data supervision, the model's estimated group memberships form clear boundaries that largely align with the true memberships, except in the null interval (other examples of output can be found in the supplementary material). In terms of segmentation performance, the model's outcomes vary based on both within- and cross-area correlations, similar to human performance. However, unlike humans, the model achieves near-perfect segmentation even when the difference between within- and cross-area correlations is minimal.

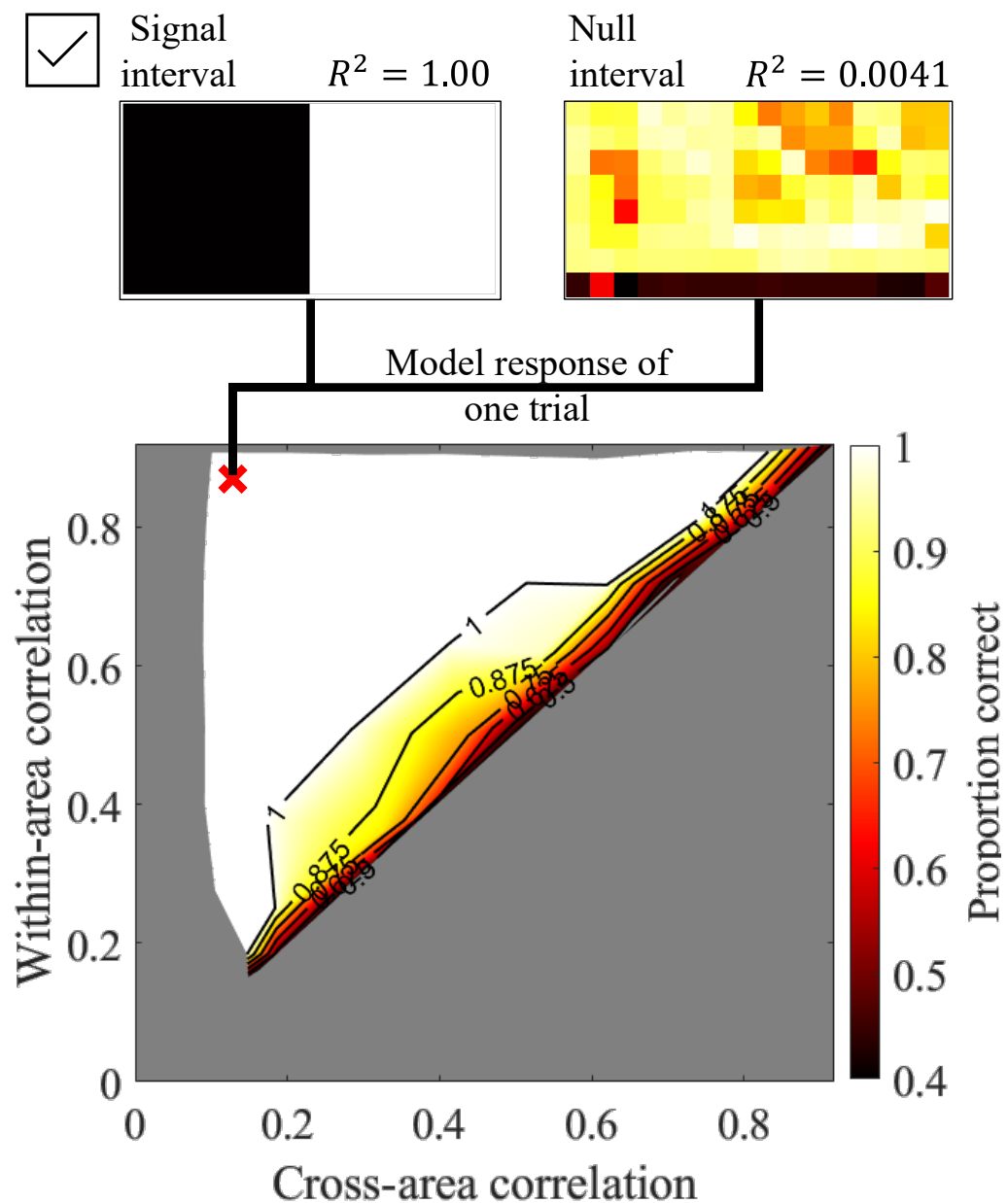


Figure 5-7. Simulation result for the naïve model.

The one-trial demo output for the second stage of the naïve model is shown at the top of the figure, with the signal interval at the left and the null interval at the right (other examples of output can be found in the 9.2). The correlation pair shown is labeled as a red cross symbol on the proportion correct heatmap.

5.4.2.2. Full model simulation results

Figure 5-8 presents the effects of the two parameters on the model's output, while Figure 5-9 presents a 2D predictability heatmap based on these parameters, along with the optimal model for three attributes. The analysis reveals that SNR is the primary factor influencing the model's overall performance, whereas the sigma parameter has

minimal impact. When the position-embedding sigma exceeds a certain threshold (e.g., greater than one element), predictability remains constant across attributes, regardless of further changes in sigma.

The optimal model achieved high predictability levels for the three attributes, with respective R^2 values of 0.96, 0.93, and 0.90. The corresponding SNR values were 0.48, 0.48, and 0.4, while the spatial embedding sigma values were 7.09, 4.73, and 7.09. These findings suggested that with appropriate parameter settings, the model can achieve performance comparable to human segmentation abilities. The lower performance in the spatial phase can be attributed to a lower SNR compared to luminance and color.

Furthermore, the position-embedding sigma primarily influenced the slope of the contour, as shown in the third row of Figure 5-8. Higher sigma values resulted in a slope closer to the diagonal line (i.e., within-area correlation = cross-area correlation). In contrast, when sigma was small, the model prioritized only short-distance correlations, causing within-area and cross-area correlations to contribute asymmetrically to segmentation performance.

The model's predictability in simulating human performance suggested that the position-embedding sigma cannot be too small, indicating that similarity across certain distant elements also plays a crucial role. This finding indicated that human visual processing incorporates not only short-range similarities, such as those between neighboring elements, but also long-range similarities across spatially distant elements.

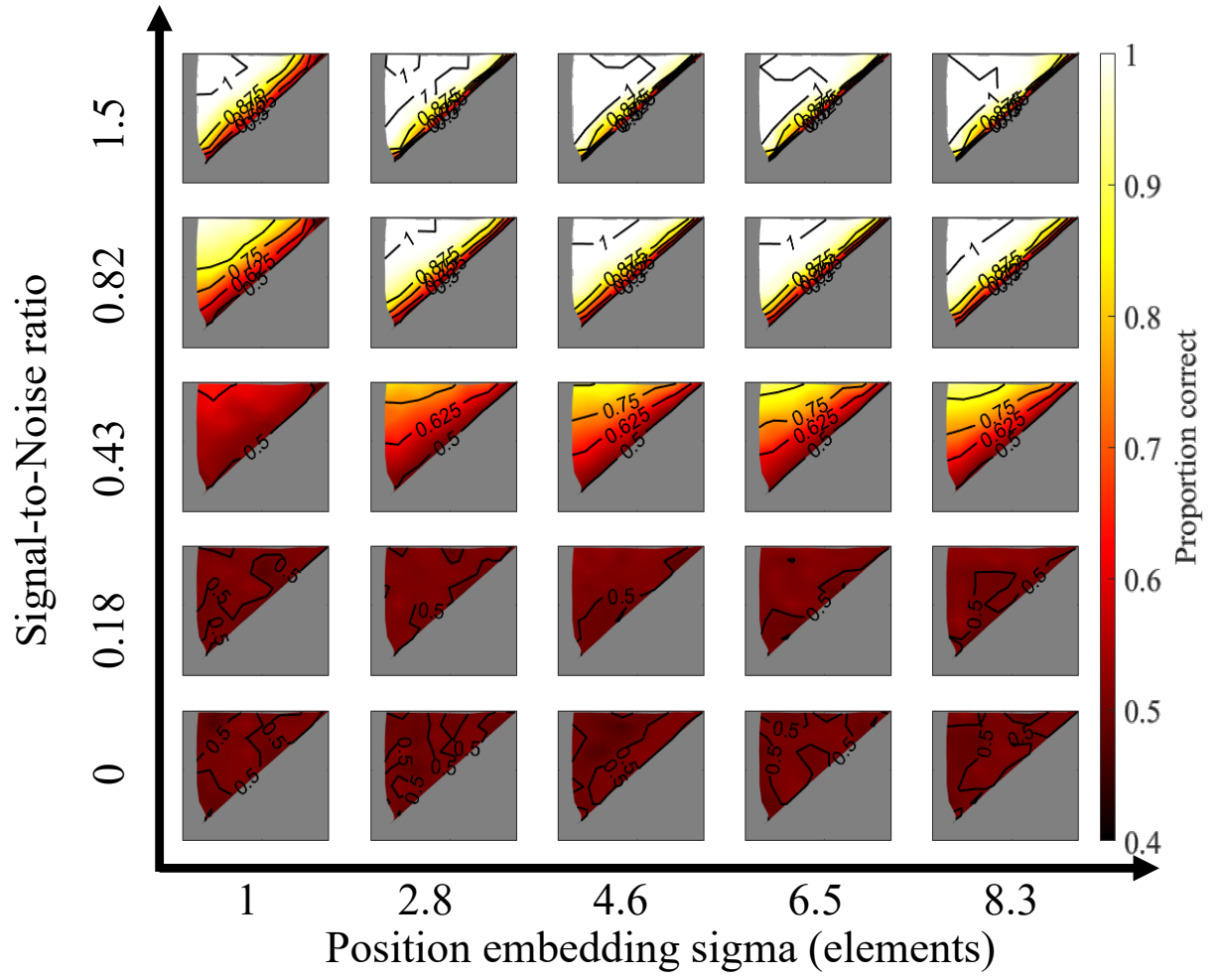


Figure 5-8. Model output of different pairs of parameters.

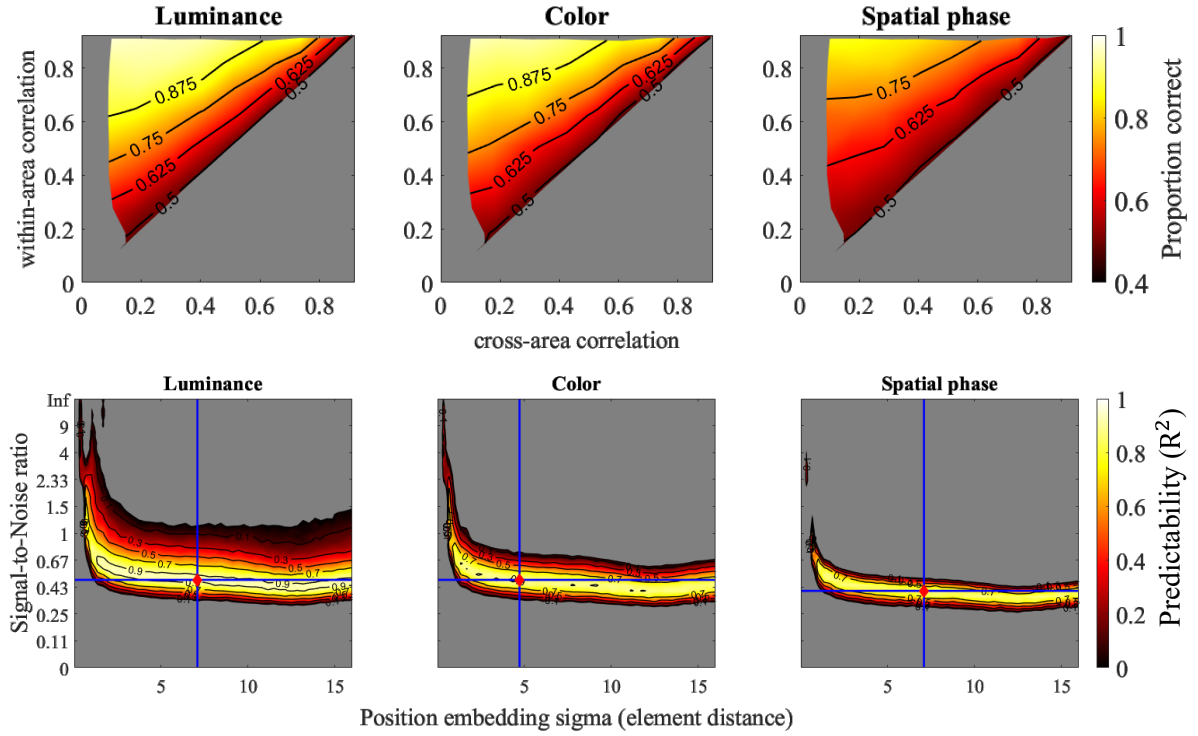


Figure 5-9. Simulation of dense parameter pairs and optimal model output.

The upper section of the figure displays the model's simulated proportion-correct values under the optimal parameter pair. The lower section presents a heatmap illustrating how well the model predicts human responses to different attributes. The optimal parameter pairs are marked as red dots, with accompanying blue vertical and horizontal lines indicating their locations on the graph.

5.5. General Discussion

This study used ViTs to generate stimuli with precisely controlled within- and cross-area correlations, allowing us to examine how the human visual system integrates similarity structures for segmentation. Our findings indicated that segmentation is not solely dependent on local comparisons between individual elements, but also involves a global process that integrates similarity across regions. Furthermore, we demonstrated that a graph-based computational model effectively captures human segmentation behavior, underscoring the potential role of structured relational processing in mid-level vision.

5.5.1. Transformer-based generation of complex stimuli

A key contribution of our work is the introduction of a novel protocol for generating temporally complex stimuli with precisely controlled element-wise temporal

synchrony. Previous studies investigating temporal synchrony have primarily relied on simple periodic waveforms, such as sinusoidal waves (Motoyoshi, 2004; Sekuler & Bennett, 2001) or square waves (Chen et al., 2024a; Kandil & Fahle, 2001; Usher & Donnelly, 1998). While some studies have explored non-periodic waveforms by randomly assigning temporal shifts to individual elements (S.-H. Lee & Blake, 1999; Morgan & Castet, 2002), these methods generally lack precise control over synchrony between individual elements.

The primary challenge in modulating element-wise temporal synchrony within a stochastic time series lies in constructing a covariance matrix that enforces a specific similarity structure. When the number of elements is large, the covariance matrix often becomes singular, making it impossible to directly sample from a statistical distribution, such as a multivariate normal distribution, while maintaining the desired correlation constraints. One potential solution is an iterative approach in which stimuli are randomly generated and tested against the desired similarity structure. However, this method is computationally expensive and inefficient, requiring an unpredictable amount of time to produce a sufficient number of trials.

In this study, we addressed these limitations by leveraging a ViT-based approach to directly optimize the temporal similarity structure. By defining a similarity matrix as a loss function, we iteratively refined the stimuli to precisely match the desired within-region and cross-region correlation constraints. This method not only ensures that the generated stimuli adhere to strict statistical properties but also provides a scalable solution for designing complex temporal textures. Our protocol offers a more flexible and computationally feasible approach for future research on spatiotemporal grouping, temporal synchrony, and other mid-level vision phenomena.

5.5.2. Graph-based representation and its role in mid-level vision

In our experiment, participants exhibited better segmentation performance when within-region correlation was high and cross-region correlation was low. This suggested that the visual system does not rely solely on individual local features, but instead integrates similarity information across multiple spatial points. Given that low-level cues (e.g., luminance, contrast, and spatial frequency) and high-level semantic information were unavailable, the mid-level visual system must rely on the relative

temporal similarity between elements to infer the segmentation structure. This finding aligned with previous studies indicating that mid-level vision constructs representations based on relational properties rather than absolute local features, such as non-accidental properties (Hummel & Biederman, 1992), configural relationships (e.g., T- and L-junctions) (Kubilius et al., 2014b), and surface inference mechanisms (Kosslyn & Osherson, 1995; Lappin & Bell, 2021).

A natural way to formalize such relational structures is through a graph-based representation. In this framework, individual elements in the stimulus can be conceptualized as nodes, while the temporal similarity between them determines the edge weights. The segmentation process then resembles graph partitioning, where regions are separated based on internal consistency and cross-boundary dissimilarity. Various methods exist for implementing graph partitioning; here, the eigendecomposition we employed is a simple linear transformation of the graph structure to extract the second-largest principal component for this task. This approach was based on previous studies demonstrating that the second-largest principal component represents the coarsest structure derivable from a graph (Fiedler, 1989; Jianbo Shi & Malik, 2000). Additionally, computational neuroscience research suggests that principal component analysis can be efficiently implemented with a small number of neurons, supporting the physiological feasibility of such computations in the brain (Oja, 1982). In our simulation, we observed that graph partitioning based on the second-largest principal component effectively replicates human segmentation performance with minimal parameterization.

Our model simulations also suggested that differences in segmentation performance across visual attributes may arise from variations in SNR. In particular, the lower segmentation performance for spatial phase indicates that higher spatial frequency signals require stronger contrast and lower temporal frequency to achieve detection levels comparable to those of low spatial frequency stimuli (Horiguchi et al., 2009; Tolhurst, 1973). Moreover, our results suggested that restricting comparisons strictly to neighboring elements, as controlled by the position-embedding parameter, does not fully account for human segmentation performance. Instead, long-range comparisons are necessary, aligning with previous studies demonstrating long-range temporal comparison mechanisms in motion perception (Maruya et al., 2013). However, interactions between temporal and spatial information are complex (Motoyoshi, 2004; Motoyoshi & Nishida, 2002) and warrant further investigation.

Alternative computational approaches could also yield results comparable to our graph-based model. For instance, while image statistics are represented in the primate cortex (Hummel & Biederman, 1992), image segmentation could be achieved by decomposing feature distributions through probabilistic inference (Dakin & Watt, 1997; Kubilius et al., 2014b). Future research should focus on identifying key stimuli and experimental paradigms to determine which algorithms are actually employed by the human visual system.

5.5.3. Comparison with CV image segmentation models

While image segmentation models in the field of CV effectively utilize spatial similarity structures for segmentation, they largely disregard the temporal aspect of the task. As a result, these models struggle with segmentation based on temporal similarity. This limitation arises because most CV models process images in a frame-by-frame manner, often neglecting the global covariation of local pixels over time. In other words, current CV approaches lack mechanisms for integrating global temporal information. In contrast, human vision relies on a global process that integrates temporal covariation signals to infer spatial structures (Alais et al., 1998; Chen et al., 2024a; S.-H. Lee & Blake, 1999; Usher & Donnelly, 1998).

At present, no mature CV models fully address segmentation based on temporal similarity. However, our study leverages generated stimuli to investigate human perception of temporal similarity and applies a graph cut algorithm to simulate a global process that selectively integrates pairwise similarities. This approach better approximates human performance compared to conventional CV models. These findings suggest that incorporating global temporal similarity processing into CV models may be a promising avenue for advancing dynamic image segmentation.

5.5.4. Relationship with current neurophysiological evidence

Neurophysiological research suggests that visual segmentation involves multiple brain regions, with specific areas recruited depending on stimulus complexity and feature composition. Numerous studies have indicated that recurrent signaling between V1, V2, V4, and the inferior temporal cortex plays a crucial role in contour integration and simple shape perception (Grossberg & Pinna, n.d.; Neumann & Sepp, 1999; Roelfsema, 2006).

For segmentation based on static visual features such as luminance, color similarity, or spatial proximity, functional magnetic resonance imaging and event-related potential studies have shown that contour-based segmentation primarily activates the lateral occipital cortex in the ventral pathway (Volberg & Greenlee, 2014; Wong et al., 2009). However, when segmentation involves more complex closed boundary shapes or object processing, activity shifts toward V3 and the intraparietal sulcus in the dorsal pathway (Seymour et al., 2008; Wu et al., 2023; Xu & Chun, 2007). The parietal lobe, particularly the inferior intraparietal sulcus, has been implicated in integrating different feature representations and processing hierarchical relationships within the visual scene (Roelfsema, 2006). In contrast, segmentation based on dynamic features, such as structure-from-motion, engages distinct regions, with the medial temporal area playing a central role (Grunewald et al., 2002).

Currently, no definitive neuronal evidence identifies the exact brain regions responsible for visual grouping or segmentation based on temporal similarity. Some studies suggest that the inferior parietal lobe and the insula contribute to processing temporal relationships, particularly temporal synchrony and asynchrony (Battelli et al., 2003, 2007; Bushara et al., 2001). This raises the possibility that temporal similarity-based segmentation relies on mechanisms distinct from those governing segmentation based on static or motion cues.

The neural basis of graph-based computation in segmentation remains an open question. Future research should explore whether the brain employs an explicit graph-based representation for visual segmentation and, if so, which regions support this computation. Given their roles in relational processing and feature binding, the intraparietal sulcus and inferior parietal lobe are potential candidates for encoding similarity structures in a graph-like manner. Integrating neuroimaging techniques with computational modeling may help elucidate the neural mechanisms underlying mid-level visual segmentation.

Chapter 6. A new way of measuring system temporal dynamics and its characteristics

Having established the overall architecture of a feature-invariant temporal segmentation system, we now turn to a more detailed examination of its temporal characteristics. While previous chapters clarified how the system integrates information across attributes and responds to temporal signals, the precise dynamics of how temporal correlation is computed remain to be quantified. Chapters 6 and 7 are dedicated to this goal. In this chapter, we focus on establishing the methodological backbone for this investigation by introducing a novel behavioral framework based on position tracking (PT) and attribute tracking (AT) tasks. This framework is designed to provide the temporal resolution and structural separation necessary for probing the internal dynamics of the segmentation system in a systematic and interpretable manner.

6.1. Introduction

The temporal processing of the visual system is crucial for survival in a dynamic world, having been extensively studied regarding temporal impulse response, temporal integration window, and so on (Watson, 1986). While several psychophysical methods have been used for this purpose, a novel approach called the continuous tracking task has been recently developed as an efficient method to estimate the temporal response of the visual system. In this task, participants need to track a constantly and quasi-randomly moving target using either mouse movements (Bonnen et al., 2015) or eye movements (J. Mulligan et al., 2013). By analyzing the tracking data, one can estimate the temporal response of the entire system from visual input to motor output. Specifically, by computing the temporal cross-correlation between the target and response trajectories, called the cross-correlogram (CCG), one can estimate the temporal kernel characterizing the system's response. For linear time-invariant system, the CCG approximates a scaled version of the system's impulse response (Ambrosi et al., 2022; Burge & Cormack, 2020). More generally, in the context of nonlinear systems, the CCG corresponds to the first-order Volterra-Wiener kernel (see Marmarelis & Marmarelis, 1978; Schetzen, 1981), providing a useful but limited characterization of the system's behavior.

The standard continuous tracking asks the participants to track the spatial position of the target stimulus; however, this task can also be extended to track stimulus changes in a dimension other than the spatial position. For example, Burge and Cormack (2020) asked the participants to estimate the depth of the target presented in a stereogram display. Ambrosi et al. (2022) varied the number of dots over time and asked the participants to move the mouse to maintain the number of dots as constant as possible. In this respect, we can divide the continuous tracking tasks into two subtypes: position tracking (PT), which requires the participant to track the spatial location of the target, and attribute tracking (AT), which requires participants to track the value of a specific attribute (e.g., luminance, color contrast). Either using PT or AT, one can estimate, from CCGs, the response of the entire system from visual input to motor output, but the relationship between the CCGs of the two types of continuous tracking tasks is not well established.

This study aimed to answer the following questions. Are the temporal dynamics revealed by CCG different between PT and AT? If so, where do the differences come from? How much are the underlying visuomotor mechanisms shared between PT and AT?

Previous studies often ascribe the differences in CCG to the temporal difference in the process encoding the sensory signals. For example, Barnett et al. (2024) introduced a target with different color contrasts in the LMS color space. They demonstrated that CCG peak latency is consistent with the temporal dynamics of the responses of L and S cones. Bonnen et al. (2017) found that the CCG of a 3D motion-tracking task showed a delay in comparison to the CCG of a 2D motion-tracking task. They suggested that the sluggishness of the 3D tracking might be caused by the detection of motion in depth taking longer than the detection of 2D motion. Therefore, when comparing the CCGs between PT and AT tasks, we attempted to reduce the differences in stimulus variations as small as possible.

We will report three experiments. The first experiment measured the CCGs of PT and AT for several types of visual attributes. We compared the CCGs of PT and AT for the same attributes, and revealed the similarities and differences in the effect of stimulus variables on the CCG between PT and AT. The second experiment attempted to identify the sensory and motor factors producing the quantitative differences in CCG between

PT and AT. The final experiment introduced a simple stimulus detection task and an attribute choice task to see the relationship between the traditional reaction time measurements and peak latencies of CCG of PT and AT tasks.

6.2. Experiment 4-1. CCG difference between PT and AT

6.2.1. Method

6.2.1.1. Apparatus

The stimuli were displayed on a VIEWPixx /3D LCD monitor (VPixx Technologies, Saint-Bruno-de-Montarville, Canada) with a 1920×1080 pixels resolution and a 120 Hz refresh rate. The lowest, mean, and highest display luminance was 1.8, 48.4, and 96.7 cd/m^2 , respectively. The maximum luminance values of the red, green, and blue primaries were measured as 27.9, 51.6, and 17.0 cd/m^2 , respectively. The corresponding CIE 1931 chromaticities of the display primaries were (0.640, 0.330) for red, (0.300, 0.600) for green, and (0.150, 0.060) for blue. The linear output of luminance was calibrated using an i1Pro chromometer (VPixx Technologies) for each channel. Each pixel subtended 1.3 arcmin at a viewing distance of 70 cm. A standard USB mouse was used to record the tracking data. The participant was seated in a dark room with a chinrest to stabilize the head. The experiments were programmed using the Psychophysics Toolbox (Brainard, 1997) in MATLAB release 2022a (Math Works, Natick, MA, USA).

6.2.1.2. Participants

Two of the authors and four naïve participants (six males, mean age 24.67 years) participated. Participants declared that they had normal or corrected-to-normal vision. They were informed of the nature of the research and provided informed consent before the start of the experiment. The naïve participants received monetary compensation after the experiment was completed. The study protocol followed the ethical standards of the Declaration of Helsinki, except for registration, and was approved by The Ethics Committee of Kyoto University (KUIS-EAR-2020-003).

6.2.1.3. Stimuli

In the position tracking task (PT), the participants were shown a horizontally moving circular target, subtending 4° in diameter, with a yellow point (0.2° in diameter) indicating the current mouse location for feedback (Figure 6-1). The target location changed in the following way:

$$P(t) = \sum_{t'=0}^t V(t') \quad (6-1)$$

$$V(t') \sim N(0, \sigma_p^2) \quad (6-2)$$

The $P(t)$ is the spatial location as the function of time (t). It is the integral over the velocity (shift per frame). The velocity of each frame was independently sampled from a Gaussian distribution, where σ_p is 0.4 min/frame. Thus, the moving trajectory was the integral of Gaussian white noise as in the Brownian motion. We generated the trajectories of 10 secs (1200 frames), and selected those consistent with the following constraints: the maximum range of the moving trajectory was 20° , which was 20 % shorter than the physical width of the monitor; the target had to shift at least 10° from the initial location to ensure that the trajectory covered a sufficiently large range for participants to produce a continuous and reproducible tracking movement. The rejection rate of the randomly generated trajectory was about 70%. Though the rejection rate was high, we confirmed that the retained trajectories preserved white noise characteristics by verifying that their average autocorrelation exhibited a sharp peak at zero lag and negligible values elsewhere.

The participants tracked three different types of targets: a 2D luminance Gaussian blob, color-contrast concentric Gabors (red-green, and blue-yellow), and luminance-contrast concentric Gabors (eight spatial frequencies). The general stimuli profile for RGB channels is as follow:

$$\begin{bmatrix} R(x, y, t) \\ G(x, y, t) \\ B(x, y, t) \end{bmatrix} = L_{mean} \cdot \left(1 + \cos(2\pi\omega D(x, y, t) - \phi) \cdot \begin{bmatrix} C_R \\ C_G \\ C_B \end{bmatrix} \right) \cdot e^{-\left(\frac{D(x, y, t)}{2\sigma_T}\right)^2} \quad (6-3)$$

$$D(x, y, t) = \sqrt{(x - P(t))^2 + y^2} \quad (6-4)$$

x and y are the spatial locations on the plane, where 0 is the center of the monitor. σ_T is the Gaussian window of the target, which was set to 1.2 degrees. ω is the spatial frequency, when the target is the 2D luminance Gaussian blob, it was set as zero cpd. When the target is the color-contrast concentric Gabors (both red-green and blue-yellow), the value of ω was set as 1 cpd. ϕ is the spatial phase, which is either 0 or 180 degrees and randomized across the trials but fixed within the trial. The vector $\begin{bmatrix} C_R \\ C_G \\ C_B \end{bmatrix}$ is the color contrast vector, when the target is the 2D luminance Gaussian blob or luminance-contrast concentric Gabors, it was set as $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$; while the target is red-green concentric Gabor, it was set as $\begin{bmatrix} 1 \\ -L_{Iso.RG} \\ 0 \end{bmatrix}$, and while the target is blue-yellow concentric Gabor, it was set as $\begin{bmatrix} -L_{Iso.BY} \\ -L_{Iso.BY} \\ 1 \end{bmatrix}$. $L_{Iso.RG}$ is the iso-luminant index between red and green while $L_{Iso.BY}$ is that between blue and yellow.

A minimum motion method was used to measure the equiluminant ratio (in terms of the system intensity value) between red (R) versus green (G) and blue (B) versus yellow (Y) for each participant before the main experiment. During the procedure, participants saw a Gabor drifting rightward. The sinusoidal waves of R and G or B and Y are 180 °out-of-phase. The participants were required to adjust the ratio between the channels to minimize their sense of motion. The mean R: G ratio value was 0.58, and the mean B: Y ratio was 0.20.

In the Attribute tracking task (AT), the participants were shown a 4 ° circular target. Instead of the spatial location, the attribute value varied in the following way:

$$A(t) = \sum_{t'=0}^t V(t') \quad (6-5)$$

$$V(t') \sim N(0, \sigma_A^2) \quad (6-6)$$

σ_A was set to the 10 % of the possible range of each attribute value per sec. For example, σ_A was 0.083%/frame for luminance since it could take the value from 0 to 100%. We used the trajectory in which the attribute value shifted at least 50 % from the initial value. The rejection rate of the randomly generated trajectory was about 80%.

(as with PT, the average autocorrelation of the retained trajectories approximated a delta function). Although the standard deviation (σ) of the random walk was defined in different units for PT and AT, participants responded using the same 20° spatial scale via mouse movement. When expressed on this common response scale, the standard deviation of displacement per frame was 0.4 and 0.498 arcmin/frame for PT and AT, respectively. Similarly, the standard deviation of the final position across trials was 3.80° for PT and 3.85° for AT. These values indicate that the overall stimulus variability experienced by participants was highly comparable between the two tasks.

Below the target, we also showed a feedback pattern of the same size as the target, whose attribute value varied according to the participants' mouse position (Figure 6-1).

While we used soft-edge patterns (Gaussian blob and concentric Gabor) for the target of PT, we used hard-edge patterns (circular patch and concentric patch) for the target of AT to avoid changes in the apparent size with the change in the attribute value. To facilitate spatial localization of the attribute-defined target and feedback, a thin black boundary (3-pixel width) was consistently added around the edge of the AT stimulus using a built-in drawing function in Psychtoolbox.

The luminance patch:

$$L(x, y, t) = L_{\text{mean}} \cdot (1 + H(x, y) \cdot A(t)) \quad (6-7)$$

$$H(x, y) = \begin{cases} 1, & D \leq 2^\circ \\ 0, & \text{otherwise} \end{cases} \quad (6-8)$$

$$D = \sqrt{x^2 + y^2} \quad (6-9)$$

The Red-Green contrast concentric patch:

$$\begin{cases} R(x, y, t) = L_{\text{mean}} \cdot (1 + H(x, y) \cdot A(t) \cdot \cos(2\pi\omega D - \phi)) \\ G(x, y, t) = L_{\text{Iso.RG}} \cdot L_{\text{mean}} \cdot (1 - H(x, y) \cdot A(t) \cdot \cos(2\pi\omega D - \phi)) \\ B(x, y, t) = L_{\text{mean}} \end{cases} \quad (6-10)$$

Blue–yellow contrast concentric patch:

$$\begin{cases} R(x, y, t) = L_{\text{mean}} \cdot (1 + H(x, y) \cdot A(t) \cdot \cos(2\pi\omega D - \phi)) \\ G(x, y, t) = L_{\text{mean}} \cdot (1 + H(x, y) \cdot A(t) \cdot \cos(2\pi\omega D - \phi)) \\ B(x, y, t) = L_{\text{Iso.BY}} \cdot L_{\text{mean}} \cdot (1 - H(x, y) \cdot A(t) \cdot \cos(2\pi\omega D - \phi)) \end{cases} \quad (6-11)$$

The luminance contrast concentric patch:

$$C(x, y, t) = L_{\text{mean}} \cdot (1 + H(x, y) \cdot A(t) \cdot \cos(2\pi\omega D - \phi)) \quad (6-12)$$

The $H(x, y)$ is a step function to define a circular area of 4-deg diameter.

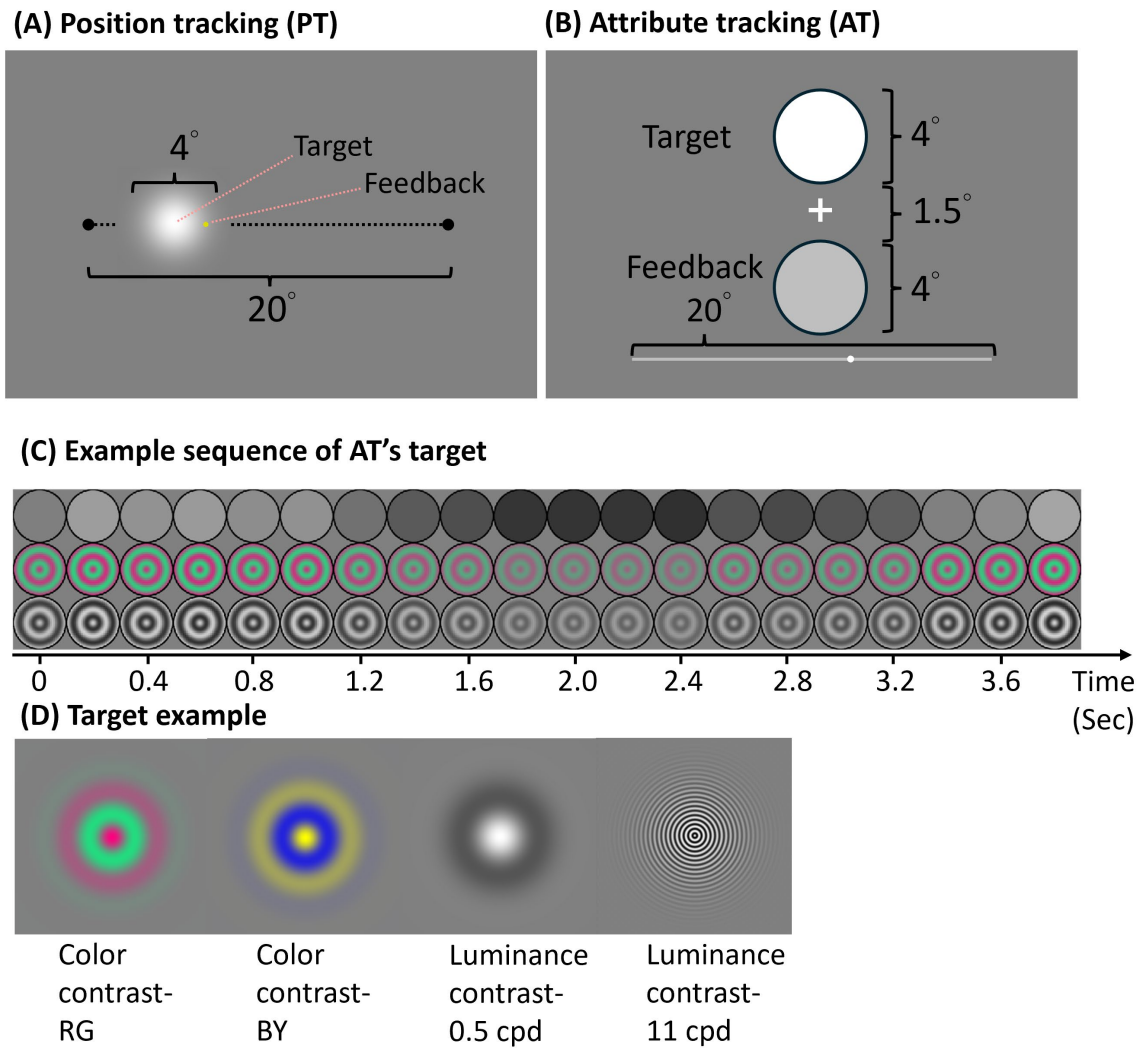


Figure 6-1. Stimuli for Experiment 4-1.

(A) Spatial arrangement of the PT task. The target is a luminance Gaussian Blob. The small yellow dot is used to provide feedback to participants about the current position of their mouse. The dotted line shows the possible shifting range; it is only for visualization here, and the participants did not see it during the experiment. (B) Spatial arrangement of the AT task. The attribute value of the reference corresponds to the participant's mouse. (C) Example sequences of AT for luminance, color contrast, and luminance contrast, respectively, from the top row to the bottom row. (D) Examples of the color contrast concentric Gabor and two types of luminance contrast concentric Gabor used for PT.

6.2.1.4. Design and Procedure

We measured the PT and AT's CCG of tracking three attribute classes: luminance, color contrast, and luminance contrast, respectively. Within each class, the number of target types was different: One in luminance, two in color contrast (red-green and blue-yellow), and eight in luminance contrast (with spatial frequency ranging from

0.5 cpd to 11 cpd in a 1.5 cpd step). During PT, participants were asked to track the center of the target blob. In contrast, during AT, the participants were asked to match the appearance between the target and the feedback as same as possible. Notably, participants were instructed that the rightward movement of the mouse would increase the attribute value (i.e., brighter in luminance, higher in contrast), while the leftward movement of the mouse would decrease the attribute value. In each condition, the participants were required to finish 30 tracking trials, each lasting 10 s. The participants completed two types of tasks $\times$ 11 different attribute types $\times$ 30 trials = 660 trials, which took approximately two hours to complete the experiment. The sequence of the conditions was randomized.

6.2.1.5. Analysis

We used the normalized cross-correlation with the time-shifting velocity between the target and response per trial and then took the average over 30 trials to compute the cross-correlogram (CCG):

$$CCG(\tau) = \frac{1}{30} \sum_{k=1}^{30} G_{R \cdot P, k}(\tau) \quad (6-13)$$

$$G_{R \cdot P, k}(\tau) = \frac{Cov_{R \cdot P, k}(\tau)}{\sqrt{Cov_{R \cdot R, k}(\tau) Cov_{P \cdot P, k}(\tau)}}, \quad (6-14)$$

$$\tau \in [-n + 1, -n + 2, \dots, n - 1]$$

$$Cov_{R \cdot P, k}(\tau) = \begin{cases} \sum_{i=1}^{n-\tau} \Delta R_{i+\tau, k} \Delta P_{i, k}, & \tau \geq 0 \\ \sum_{i=1}^{n+\tau} \Delta R_{i, k} \Delta P_{i-\tau, k}, & \tau < 0 \end{cases} \quad (6-15)$$

Here, R represents the participant's response in a single trial and P represents the target trajectory. The differentiation operator $\Delta(X) = X_i - X_{i-1}$ captures the instantaneous velocity of the response and the target. The first 2 seconds of the data were removed to exclude unstable movements. For each condition of each participant, we had 30 $G_{R \cdot P, k}(\tau)$ and one CCG. We used the bootstrapping to estimate the standard error of CCG. Specifically, at each iteration of the bootstrapping, we resampled 30 $G_{R \cdot P, k}(\tau)$ with replacement, and then calculated the CCG. The bootstrapping process was repeated for 2500 times of iteration (i.e., $CCG_1, CCG_2, \dots, CCG_{2500}$). The extraction

of curve properties was based on bootstrapped CCGs. The normalization operation follows Mulligan (1997) to ensure the correct estimation of the coefficients of correlation for each frame. Since the current study mainly focused on the processing delay of the visual system, we analyzed the peak latency (τ^*), defined as follows.

$$\tau^* \rightarrow \underset{\tau}{\operatorname{argmax}}(\operatorname{CCG}) \quad (6-16)$$

6.2.2. Result

The individual CCGs are shown in Figure 6-2, and the statistics of the peak latency averaged across participants are shown in Table 6-1. We observed that AT exhibits a lower peak amplitude, a lower correlation coefficient, and a delayed peak time relative to the PT. This tendency was consistent across the participants and attribute types. On average, the AT's peak latency was 2.48 times longer than that of PT (ranging from 1.5 to 4.0).

In addition to the task difference (PT vs AT), the stimulus attribute influenced the CCG. The peak latency was longer for color contrast than for luminance, and an increase in spatial frequency increased peak latency. The observed trend in the attribute effect was consistent between PT and AT. A significant linear correlation existed between the peak latency of PT and AT across attribute types ($r = 0.63$, $p < 0.001$, Figure 6-3), and the linear slope ($\beta = 5.33$) of PT predicting AT significantly differed from 1 ($p < .001$).

These findings suggest that although AT exhibits slower responses compared to PT, both tasks demonstrate qualitatively similar tendencies in terms of how peak latency covaries with stimulus attributes.

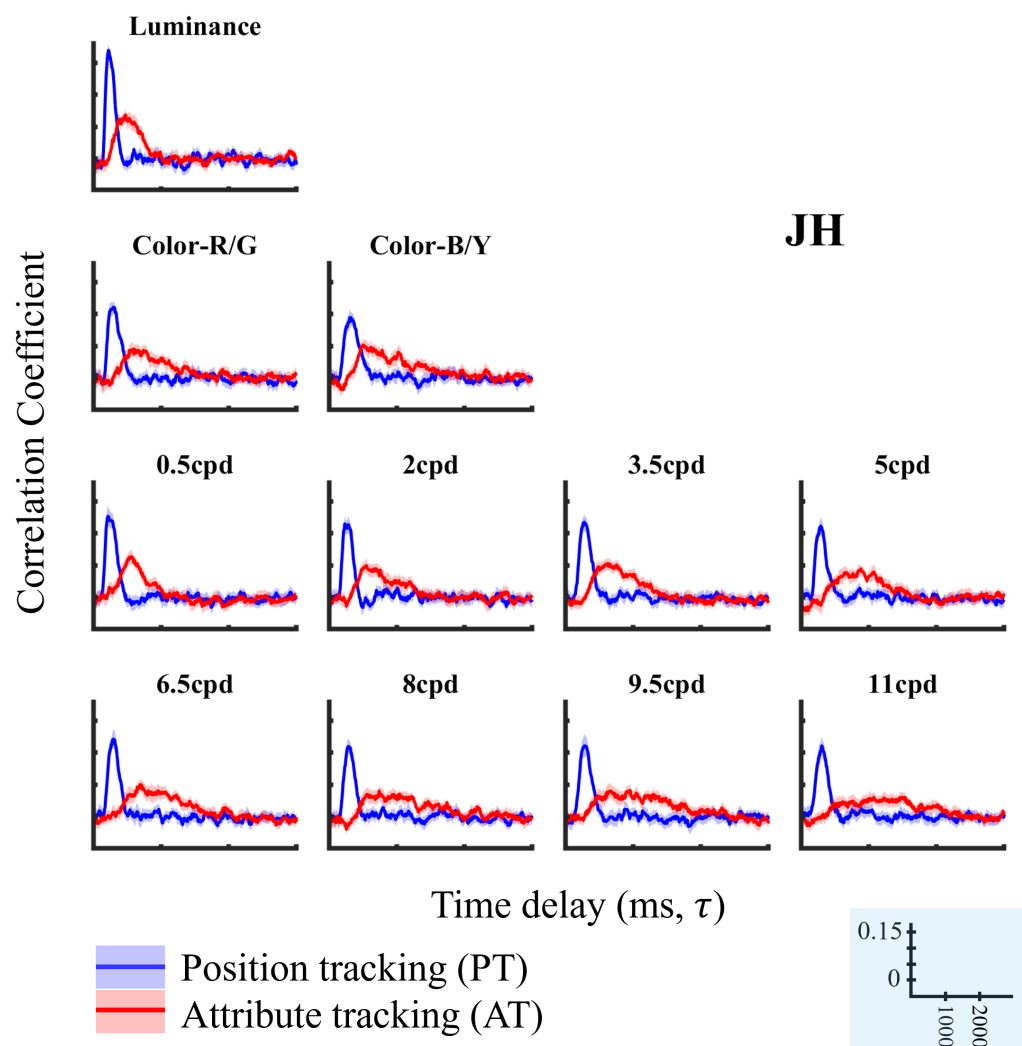


Figure 6-2 One participant's CCGs for PT and AT across different attribute types. The blue line represents position tracking and the red line represents attribute tracking. The shaded area represents the confidence interval between the 2.5 quantiles and 97.5 quantiles of the bootstrapped CCGs. The first row of the graph array represents the luminance. The second row shows the color contrast, the left shows the red-green contrast, and the right shows the blue-yellow contrast. The third and fourth rows are for luminance contrasts with different spatial frequencies, which are labeled in the title. The other participants' figures can be found in the supplementary material (9.3)

Table 6-1. Peak latency statistics between PT and AT across attribute types.

Attribute type		Lumi- nance	Color contrast		Luminance contrast							
		Lumi- nance	Red- Green	Blue- Yellow	0.5 cpd	2 cpd	3.5 cpd	5 cpd	6.5 cpd	8 cpd	9.5 cpd	11 cpd
PT	Mean	262.72	284.54	299.80	249.48	276.34	284.27	297.31	292.75	292.02	294.78	294.02
	2.5 quantile	246.67	270.00	291.67	241.67	259.72	275.00	286.67	283.33	283.33	281.67	281.67
	97.5 quantile	270.83	295.83	308.33	261.11	291.01	297.22	305.56	302.78	302.78	305.00	305.56
AT	Mean	471.39	716.60	738.15	586.14	566.99	650.82	718.68	726.95	834.21	863.91	917.93
	2.5 quantile	425.00	645.83	680.56	490.28	529.17	576.86	622.22	635.20	719.44	729.39	770.83
	97.5 quantile	530.56	785.00	802.78	762.10	623.61	738.89	819.74	825.00	937.50	991.01	1047.22

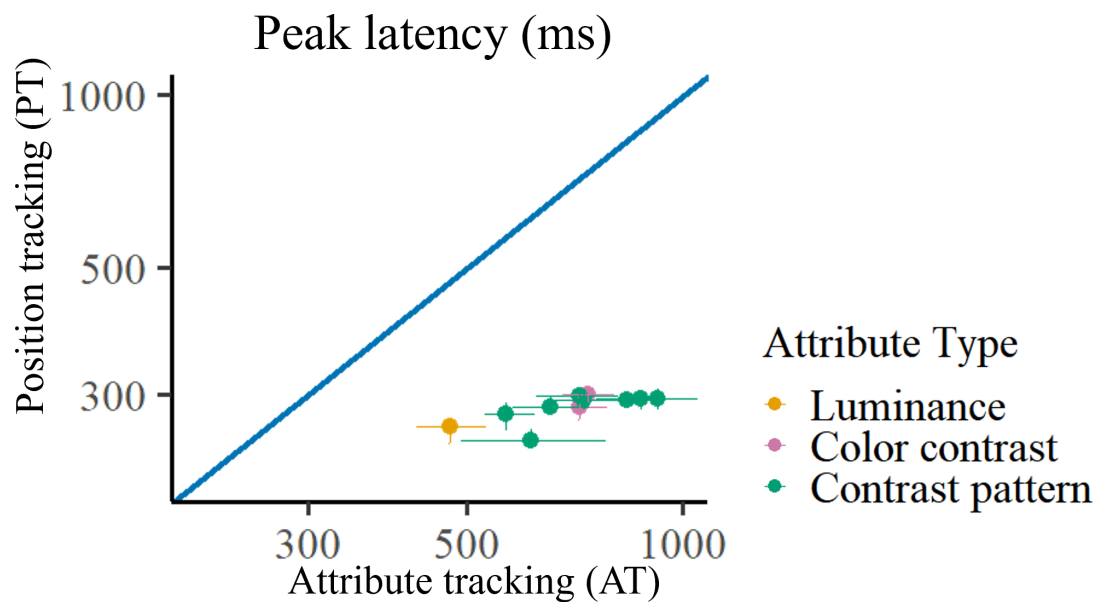


Figure 6-3. Correlation Between PT and AT.

(A) Scatter plot of PT and AT across attribute types. The three colors represent luminance, color contrast, and luminance contrast, respectively. The vertical and horizontal lines attached to the dots are the confidence interval between 2.5 quantiles and 97.5 quantiles of the bootstrapped peak latency. The blue diagonal line is the reference line, with a slope of 1.

6.2.3. Interim discussion

The first experiment evaluated the CCG for various visual stimuli and examined the gap between PT and AT. Our findings of the effects of stimulus characteristics on the human visual temporal responses were consistent with previous reports using non-tracking tasks. First, the peak latencies were longer for color contrast than for luminance. This echoes the consensus that the processing of chromatic stimuli is slower than that of achromatic stimuli, as shown by psychophysics (Burr & Morrone, 1993; Uchikawa & Yoshizawa, 1993; for a review, Watson, 1986), or exhibit frequency tuning toward lower temporal frequencies, as shown by vision evoked potentials (VEPs) and functional magnetic resonance imaging (fMRI) studies (D'Souza et al., 2011; Kulikowski et al., 2002; Mullen et al., 2010). Second, we observed a gradual increase in peak latency with an increase in spatial frequency. This observation replicated previous psychophysical studies showing that higher spatial frequencies require longer integration times to detect and recognize (Ejima & Ohtani, 1987; Harris & Georgeson, 1986; Kelly, 1979; Legge, 1978).

Our findings revealed that tracking the AT value took longer than tracking the target location (PT), but both tasks had a qualitatively similar pattern regarding how response latencies change with the attribute. This observation suggests that PT and AT share a part of the processing from vision to action, at least where the stimulus attribute influences the response speed, although the effect of the stimulus attribute was weaker, or more saturated, for PT than for AT. What mechanisms produced the differences between PT and AT?

Figure 6-4 illustrates a putative information-processing pipeline for the PT and AT tasks. The first stage is the low-level perceptual process used to encode the attribute and its change. The two tracking tasks can share this initial stage but generate different stimulus representations for an appropriate action. Specifically, PT requires the system to estimate the position of the target, while AT requires a system to estimate the value of the target. Then, the processing system translates each of the internal representations of the estimated values into a mouse movement.

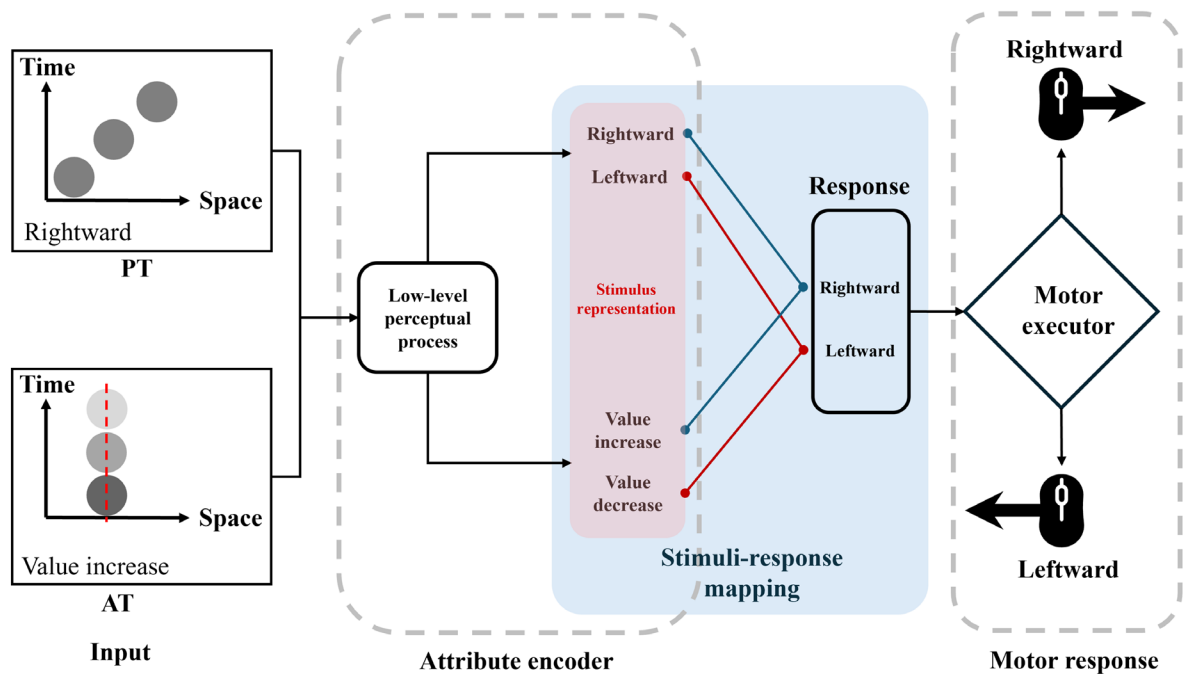


Figure 6-4. putative information-processing pipeline for the PT and AT tasks.

Both tasks require the same low-level process to encode the variation in the attribute (e.g., luminance change) and read out different aspects; the target movement is estimated during PT, while the target stimulus content is estimated during AT. After the attribute encoder stage, the system must determine how to relate the stimulus representation with the task-relevant response (stimulus–response mapping). The positional movement of the target is mapped to the corresponding hand movement during PT, while the increase/decrease in the target is mapped to the rightward/leftward hand movement during AT. Finally, the hand movement command realizes the mouse movement.

In this framework, we identified two potential sources of AT sluggishness. First, mapping is more direct in PT during the stimulus–response (S–R) mapping stage. A rightward stimulus motion is mapped to a rightward hand movement, while a leftward stimulus motion is mapped to a leftward hand movement. On the other hand, an increase in the value is arbitrarily mapped to a rightward hand movement in AT, while a decrease in the value is arbitrarily mapped to a leftward hand movement. Research on reaching limb actions shows that an arm or limb movement is automatically or involuntarily driven by target motion (Pisella et al., 2000), or by large background motion (Gomi, 2008), in the same direction as visual motion with a very short latency. A mouse's movement is likely assisted by a similar mechanism during PT. In contrast, there is no such visual assistance for the hand or limb to react to changes in luminance or color in AT. Thus, the longer latency for AT than PT may be rooted in the voluntary S–R mapping arbitrarily determined by the experimenter.

Second, the local change in magnitude was generally larger during the attribute encoding stage in PT than in AT. In Experiment 4-1, the change in intensity was

approximately 2.4 times higher for PT than AT when comparing the mean differences in peak intensity between successive frames. The smaller the magnitude of the local intensity variation, the slower the response of the visual system, arguably due to the increase in perceptual uncertainty. This phenomenon has been demonstrated for luminance (Hess effect; Williams & Lit, 1983), color contrast (O'Donnell & Colombo, 2008), and luminance contrast (Ejima & Ohtani, 1989). Thus, AT can be more sluggish than PT because of smaller changes in magnitude. In addition, if the effects of stimulus variables on PT are saturated by high stimulus change intensities for PT, the reduction in stimulus change values will make the effects of stimulus variables more visible.

Accordingly, in Experiment 4-2, we investigated how voluntary S–R mapping and lowering local variation in the magnitude could change the peak latency of the PT task. If the sluggishness of AT could be reproduced by PT with voluntary mapping and low stimulus intensity, we could infer that they are the factors driving the difference between PT and AT.

6.3. Experiment 4-2. S–R mapping and the effect of stimulus intensity on position tracking

In Experiment 4-2, we modified the original PT task with S–R mapping and stimulus intensity. To make PT S–R mapping voluntary, we changed the motor task to "antireaction" (Saijo et al., 2005). In the antireaction task, S–R mapping was reversed so that a rightward target movement was mapped to a leftward hand movement, while a leftward target movement was mapped to a rightward hand movement. Note that in our antireaction PT, we provided feedback on the mouse position in the same way as in the original PT condition: i.e., right-to-right and left-to-left without response reversals. Therefore, participants had to learn a new way of responding but not a new visuomotor control skill.

To reduce the magnitude of the stimulus intensity during PT, we reduced the peak intensity of the target to 1/10 of that under the original PT conditions in Experiment 4-1. We assumed that this intensity was sufficiently large to counterbalance the 2.4-fold difference in stimulus intensity between PT and AT in Experiment 4-1, even after considering the nonlinearity (e.g., local adaptation, contrast gain control) of the visual response.

Experiment 4-2 involved three tasks, such as PT_Inv, an antireaction version of PT, where only S–R mapping was reversed; PT_LV, a low-value version of PT where the target stimulus intensity was reduced to 10%, and PT_InvLV, where both target magnitude and the S–R mapping were manipulated. Other than these modifications, the stimulus, procedure, and analytical protocols were identical to those used in Experiment 4-1. The data of the same six participants were collected, which took about 3 hours per participant.

6.3.1. Result

Visual inspection of the obtained CCGs for each participant and stimulus condition (Figure 6-5) suggests that both antireaction mapping and stimulus intensity reduction contributed to delaying the response latency and reducing the peak amplitude of PT, and making the CCG close to that of AT, although the magnitude of the effects varied depending on the stimulus attribute condition.

Figure 6-6 compares, between four PT conditions (including the original PT in Experiment 4-1) and AT, the peak latencies of different attribute conditions averaged across participants. To evaluate the similarity between PT and AT, we compared the Pearson correlation, the slope of the regression line, and the average offset from the identity line. For PT_Inv, while the mean offset was reduced compared to PT, the correlation and slope were nearly zero. For PT_LV and PT_InvLV, the correlation was high ($r = 0.76, 0.73, p < .001$), and the slope ($\beta = 0.71, 0.859$) was not significantly different from the identity line (i.e., $PT = AT, p = 0.076, 0.456$), but PT_InvLV exhibits a smaller offset from the identity line compared to PT_LV (-106 ms vs. -284, $p < .001$).

The findings suggest that decreasing the stimulus intensity enhances the latency difference among stimulus attribute conditions and that voluntary mapping adds a constant increase in latency. Furthermore, the combination of the two factors makes the PT's temporal response fairly close to the AT's temporal response. The results suggest that stimulus magnitude and stimulus-response mapping are two major factors that produce the differences between PT and AT.

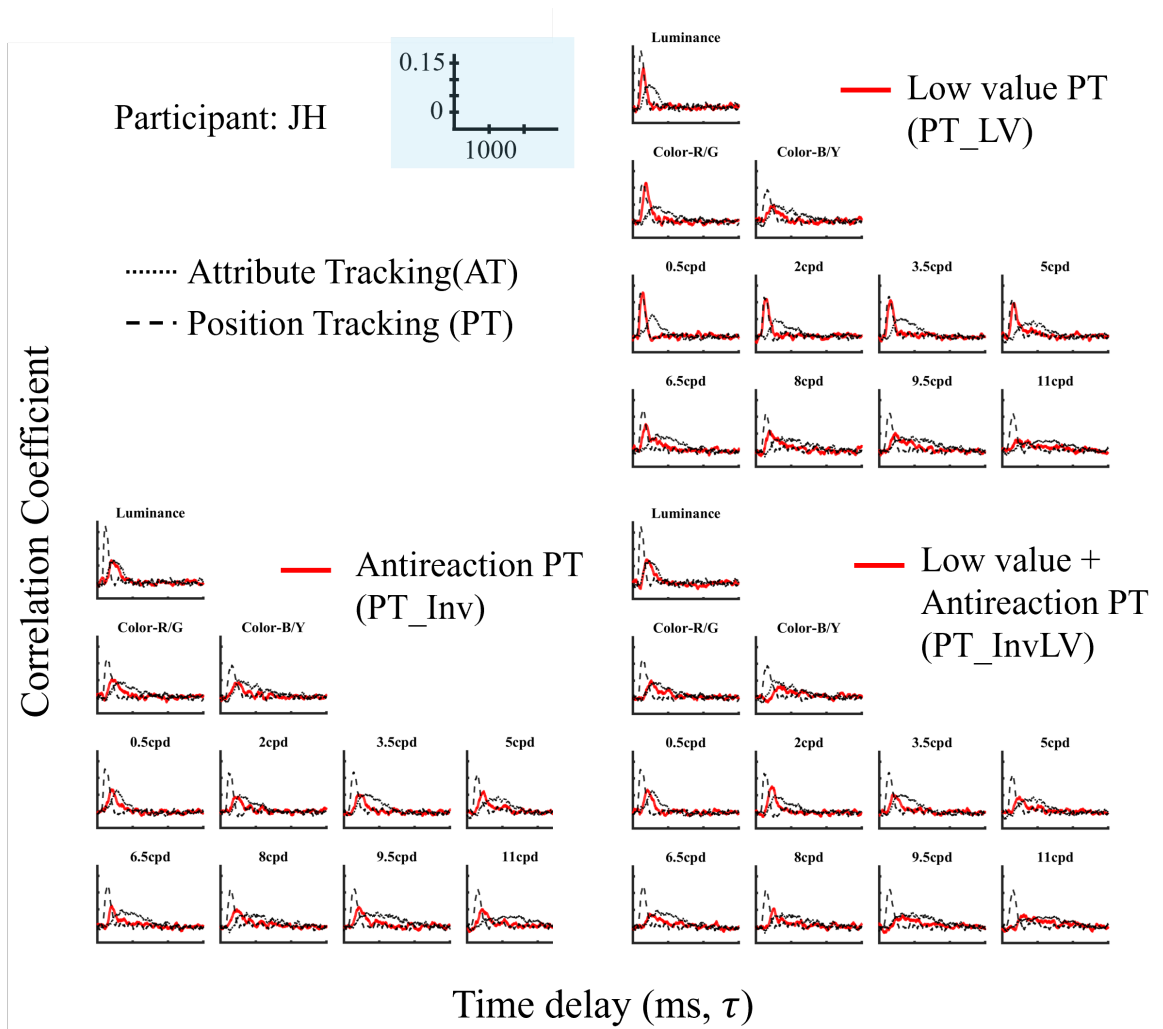


Figure 6-5 The CCG of one participant across three PT modulations and attribute conditions.

The CCG of one participant across three PT modulations and attribute conditions. A dotted line represents the AT results, while a dashed line indicates PT from Experiment 4-1. The solid red line denotes various PT variations, which are labeled in each subgraph. The 95% confidence interval (between the 2.5 and 97.5 quantiles estimated by bootstrapping) is illustrated by the shaded region. The first row of the graph array illustrates the luminance. The second row shows the color contrast, the left graph shows the red-green contrast, and the right graph shows the blue-yellow contrast. The third and fourth rows depict the luminance contrasts with varying spatial frequencies. The other participants' figures can be found in the appendix (9.3).

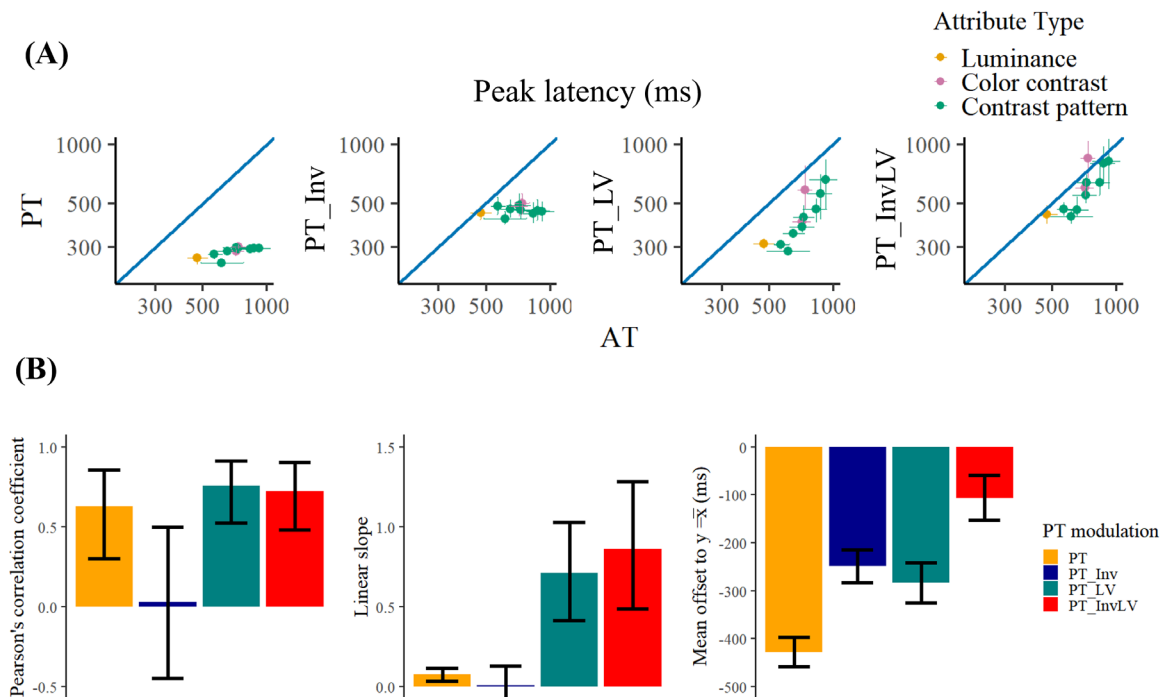


Figure 6-6. scatter plot and the statistics across conditions.

(A) The peak latency averaged across participants. Against AT were plotted four conditions of PT, from left to right: original (Experiment 4-1), antireaction (Inv), low stimulus intensity value (LV), and both modulations (InvLV). Different color symbols represent different attribute types. The error bar associated with each dot indicates the 95 % confidence interval. The black diagonal line represents an identity line, where the slope equals one, and the intercept equals zero. (B) The measures of similarity in the peak latency between AT and different variants of PT. The graph from left to right shows the Pearson correlation, linear slope, and mean offset, respectively. The error bars associated with each bar indicate the 95% confidence interval. Note that Correlation=1, Slope=1, and Offset=0, when Latency(PT)=Latency(AT).

Since inverse tracking introduces a novel motor task for participants, we investigated whether the latter trials exhibited a learning effect compared to the earlier trials. To quantify the learning effect, we computed the latency change rate by subtracting the median peak latency of the last 15 trials from that of the first 15 trials. We then compared the learning effects across three different tracking conditions: normal, inverse, and attribute. The normal condition includes standard tracking tasks (PT and PT_LV), the inverse condition includes inverse tracking tasks (PT_Inv and PT_Inv_LV), and the attribute condition corresponds to AT. The results, shown in Figure 6-7, indicate that no latency change rate was significantly different from 0% (Bonferroni-corrected, $p > .0015$). This suggests that there was no observable learning effect across trials, regardless of the S-R mapping participants were required to use.

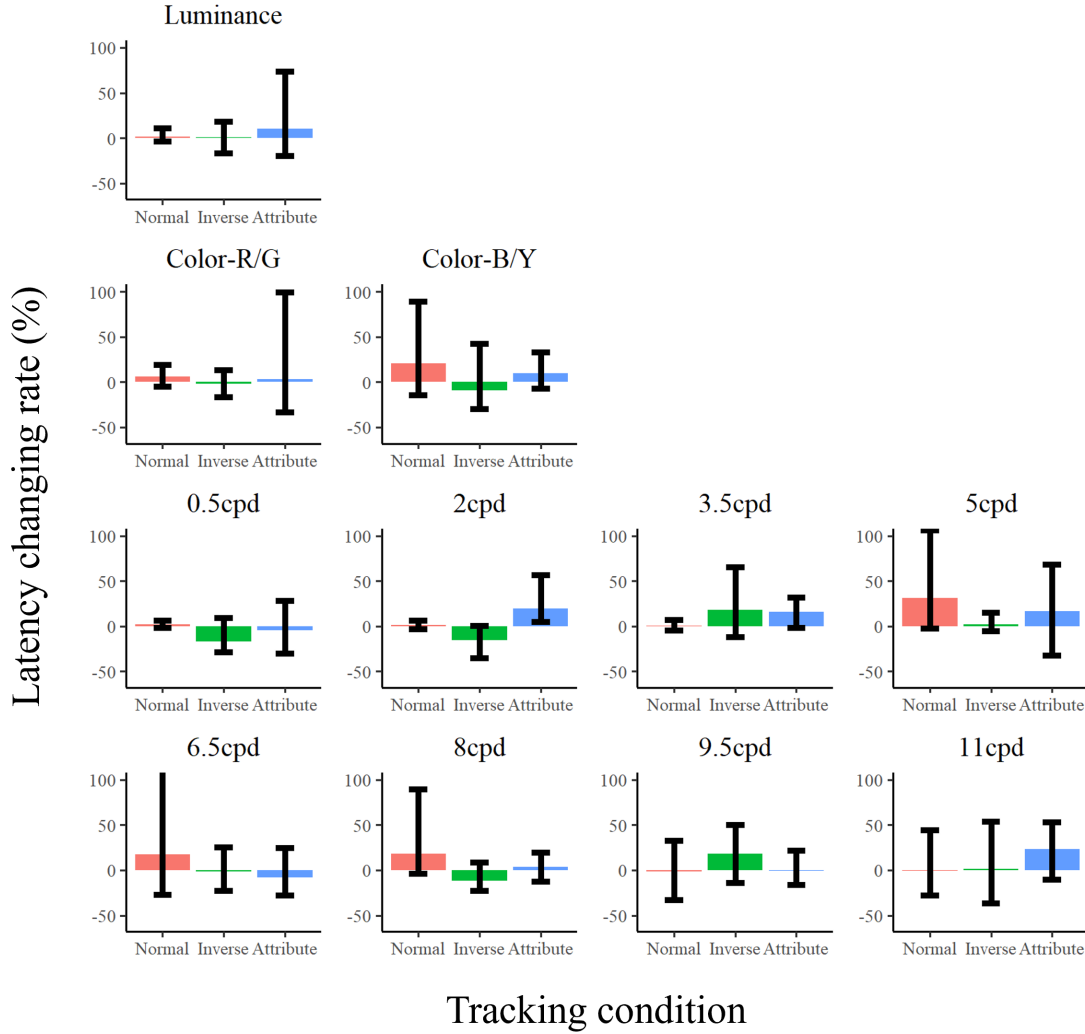


Figure 6-7. latency changing rate from the first 15 trials to the latter 15 trials across various tracking conditions and attribute types.

The horizontal axis indicates three different tracking tasks: Normal, corresponding to normal S-R mapping, stands for PT and PT_LV condition, while inverse, corresponding to inverse S-R mapping, stands for PT_Inv and PT_Inv_LV. Attribute corresponds to AT. The vertical axis is the Latency changing rate computed by the change percentage of the first 15 trials' peak latency median to the latter 15 trials'. The error bar attached to each bar is the confident interval with Bonferroni correction (i.e. 99.92 % C.I.).

Finally, we introduced a simple linear model to decompose the observed peak latency into two components: one from the attribute-specific encoding stage and the other from the stimulus–response (S-R) mapping stage. This follows the conceptual framework proposed by Brueggemann, (2007) who suggested that perceptual encoding and motor execution latencies can be approximated as linearly additive (i.e., cascaded). The model is expressed as:

$$PL(\text{Attribute}, \text{magnitude}, \text{SR}) = C_{\text{magnitude}} \cdot E_{\text{Attribute}} + M_{\text{SR}} \quad (6-17)$$

The first term $C_{magnitude} \in \mathbb{R}^2$ captures how reductions in attribute magnitude (via C_{low} for low magnitude PT, and C_{AT} for AT, otherwise set as 1) lead to increased encoding latency. $E_{attribute} \in \mathbb{R}^{11}$ represent the baseline encoding time for each stimulus type: $E_{luminance}$ for luminance; E_{color} as a 2-element vector for red-green and blue-yellow contrasts; $E_{contrast}$ as an 8-element vector for spatial frequency contrast patterns at eight levels. The final term $M_{SR} \in \mathbb{R}^3$ denotes the stimulus-response mapping component: M_{normal} for standard tracking, $M_{inverse}$ for inverse tracking, and M_{AT} for attribute tracking. This component is added uniformly regardless of the attribute type. Note that it may also reflect the latency of perceptual encoding stages that are invariant to stimulus attributes, in addition to the S–R mapping latency. We estimated the model parameters—consisting of 16 values in total—using least squares estimation (LSE) on participant-averaged data. This formulation provides a compact approximation of how stimulus properties and mapping conditions contribute to the observed latency patterns.

The modeling results revealed the following parameter estimates (with the estimate error in the parentheses): $C_{low} = 11.36 (\pm 1.20)$, indicating that reductions in stimulus magnitude substantially increased encoding latency in a stimulus dependent way, with larger increments for, e.g., high spatial frequency stimuli than for low spatial frequency ones; $E_{luminance} = 0 \text{ ms} (\pm 1.45)$, $E_{color} = [13.37(\pm 1.69), 27.44(\pm 3.07)] \text{ ms}$ for red-green and blue-yellow stimuli respectively; and $E_{contrast} = [1.20(\pm 1.49), 2.14(\pm 1.43), 5.68(\pm 1.40), 11.11(\pm 1.55), 15.43(\pm 1.84), 19.69(\pm 2.11), 28.96(\pm 3.06), 33.83(\pm 3.56)]$ across eight levels of spatial frequency.

Additionally, the magnitude-related encoding latency for AT, captured by $C_{AT} = 10.24(\pm 1.63)$ suggests that the effective stimulus strength in AT was relatively low compared to PT, consistent with our empirical observations. The S–R mapping components were estimated as $M_{normal} = 259.70 (\pm 6.33) \text{ ms}$, $M_{inverse} = 464.69(\pm 6.33) \text{ ms}$, and $M_{AT} = 565.39 (\pm 15.26) \text{ ms}$, reflecting a substantial latency increase under inverse mapping. Notably, the S–R mapping latency for AT was comparable to that of the antireaction condition, supporting the interpretation that AT involves a similarly effortful or indirect mapping process.

The model achieved a high degree of fit to the empirical data (Figure 6-8), with a coefficient of determination $R^2 = 0.97$, demonstrating that this simple linear formulation effectively captures the major components contributing to peak latency. These results suggest that both attribute-specific encoding time and mapping complexity can be dissociated and quantitatively accounted for in continuous tracking performance.

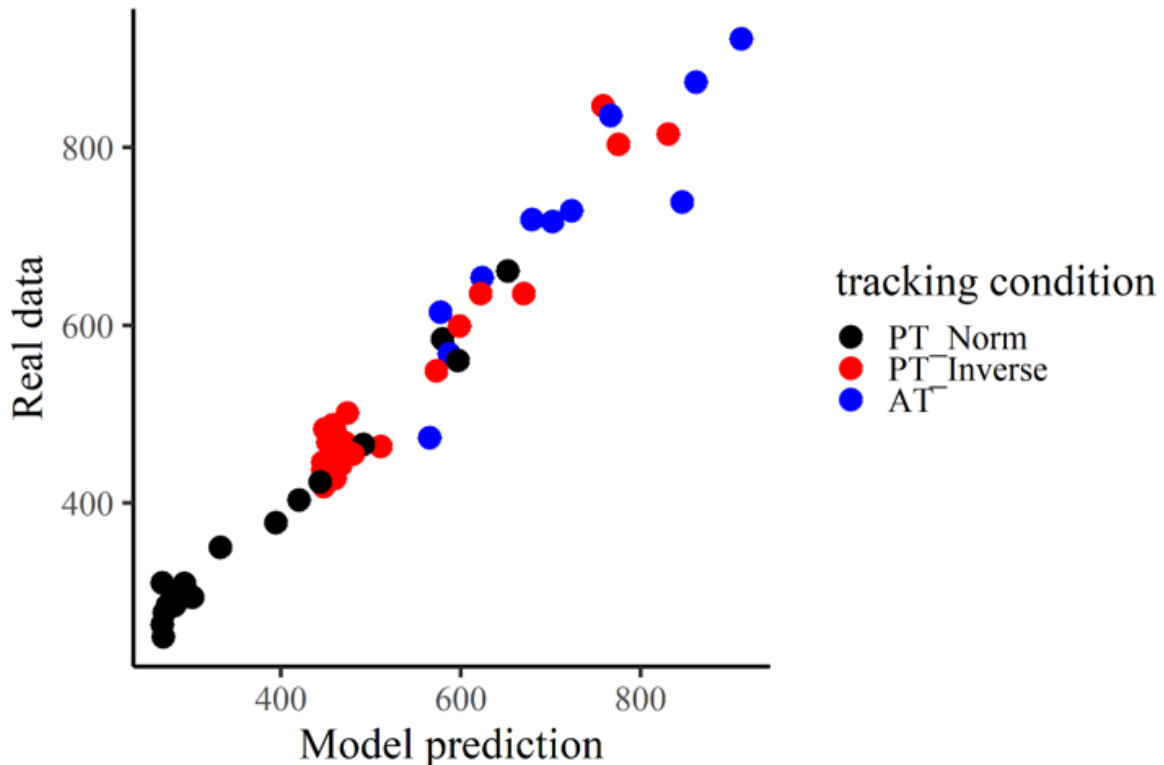


Figure 6-8. Model fit for peak latency data across all experimental conditions. The x-axis represents the predicted peak latency from the linear model, and the y-axis shows the corresponding empirical (observed) peak latency. Each point corresponds to one condition in the experiment. The colors of the dots indicate the type of tracking condition: black for position tracking with normal stimulus–response (S-R) mapping (PT_Norm), red for position tracking with inverse S-R mapping (PT_Inverse), and blue for attribute tracking (AT).

6.4. Experiment 4-3. S-R mapping and local contrast influence on reaction time

In Experiment 4-3, we investigated the relationship between the two continuous tracking tasks and traditional single-trial reaction time measurements (RT). Three different reaction time paradigms were introduced: (1) simple reaction time task, (2) choice reaction time task, and (3) choice reaction time task with a lower stimulus change magnitude. A procedure flow can be found in Figure 6-9.

This experiment had two primary objectives. The first was to examine how the peak latencies of PT and AT correlated with the two RT, aiming to gain deeper insights

into the mechanisms underlying tracking tasks, given that RT interpretation is extensively studied and well-established. Specifically, since stimulus-response mapping is more direct for simple RT than for choice RT, we expected a similarity between the relationship between PT and AT, and the relationship between simple RT and choice RT. The second objective was to establish a connection between the current continuous tracking method and RT, providing a benchmark for future research to compare results across different experimental designs.

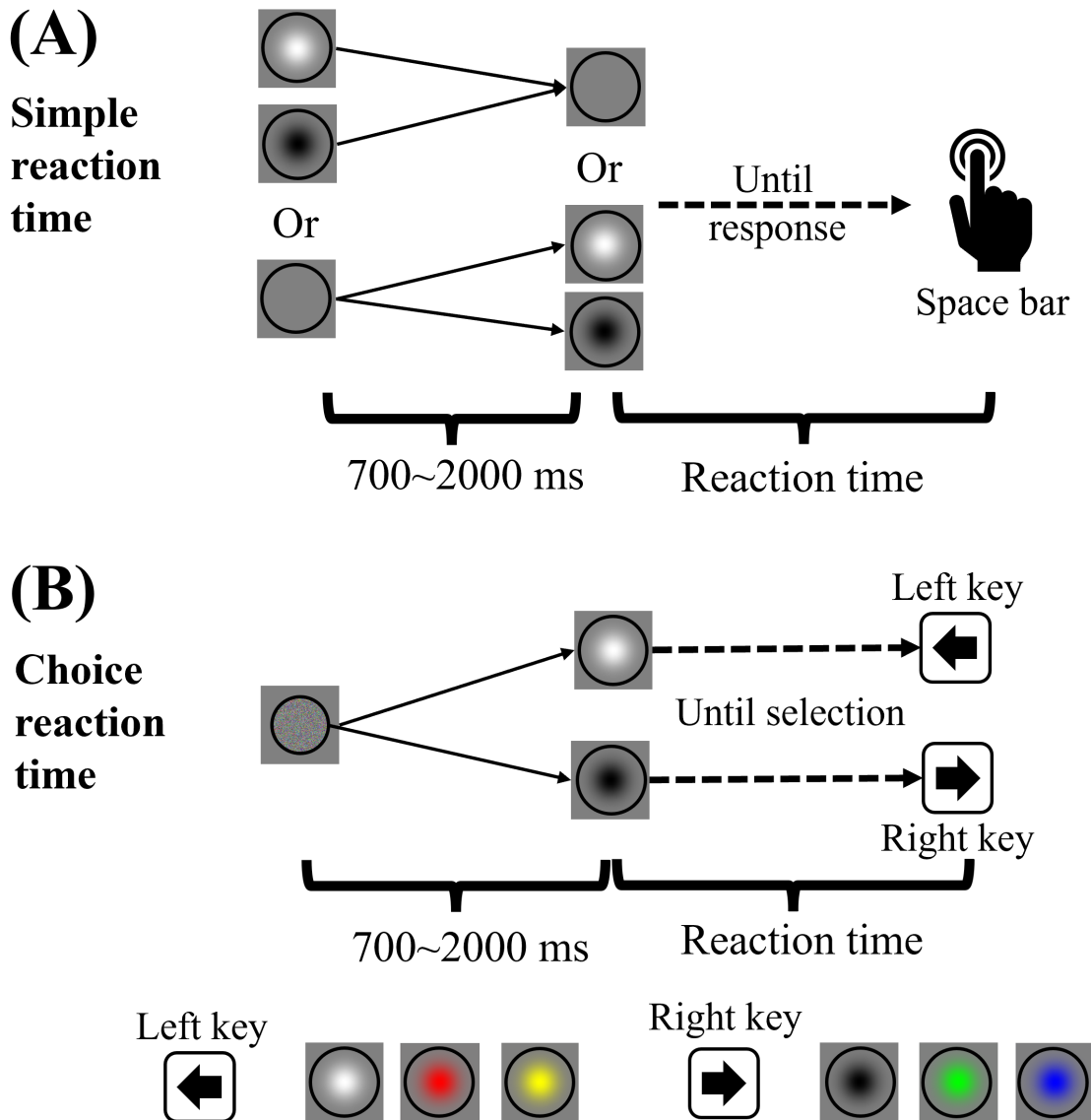


Figure 6-9. The two reaction time paradigms workflow.

(A) Simple reaction time task procedure: Initially, the stimulus was presented with or without a pattern, followed by a random interval drawn from a uniform distribution between 700 and 2000 ms, after which the original stimulus alternated. Participants were required to respond to the alternation as quickly as possible. (B) Choice reaction time task procedure: Initially, the stimulus was a white noise patch with three color channels, followed by a random interval after which the target stimulus appeared. Participants were required to determine which stimulus appeared and respond using the corresponding key. The corresponding stimulus-response pairs are shown below the procedure diagram, where the left key represents white/red/yellow and the right key represents black/green/blue.

6.4.1. Methodology

In the simple reaction time task, the stimulus was initially presented either with or without a pattern, followed by a random interval between 700 and 2000 ms, and a change in the stimuli occurred. Participants were instructed to detect the pattern change, either from visible to invisible or from invisible to visible, as rapidly as possible. The

stimulus size and the number of stimulus attribute types were consistent with those in the original PT task.

In the choice reaction time task, a colored white noise was initially presented, and after a random interval, a target stimulus was displayed. Specifically, we modify the stimulus in the color choice task from a concentric Gabor pattern to a single-color Gaussian blob, requiring participants to identify the color itself rather than detect color contrast. Participants were required to identify the type of stimulus that appeared by pressing the corresponding key, with the left key representing white/red/yellow, and the right key representing black/green/blue. The luminance contrast stimuli were not examined in the choice reaction time task because low- and high-contrast options in the choice reaction time task became similar to detectable and undetectable options in the simple reaction time task.

The choice reaction time task with a low-intensity magnitude followed the same procedure and utilized the same attribute types as the choice reaction time task, with the exception that the maximum magnitude of the stimulus was reduced to 10 % of the maximum, consistent with the settings in PT_LV.

For each stimulus and task condition, 30 trials were run. Data from the same six participants were collected in an additional hour.

6.4.2. Result

As shown in Figure 6-10, the peak latencies of PT and AT have high correlations with RT measurements in general. However, PT and AT are better aligned with different types of RT.

The PT's peak latency is aligned well with the simple reaction time, in terms of the slope of the RT as the function of peak latency ($\beta = 0.67$), and the mean offset as small as -22.5 ms. On the other side, the AT's peak latency is better aligned with the choice reaction time ($\beta = 0.3$, offset=-174ms) and best with the lower intensity version ($\beta = 0.4$, offset=-117ms).

The results demonstrated that the latency of tracking tasks aligns with that of the traditional reaction task. Specifically, the high similarity between PT's peak latency and simple reaction time suggests that PT's latency primarily reflects the processing

time required for detecting the occurrence of stimulus change. On the other hand, the high similarity between AT's peak latency and simple reaction time suggests that AT's latency reflects not only the detection of stimulus change but also the identification of the stimuli content and the selection of the corresponding response.

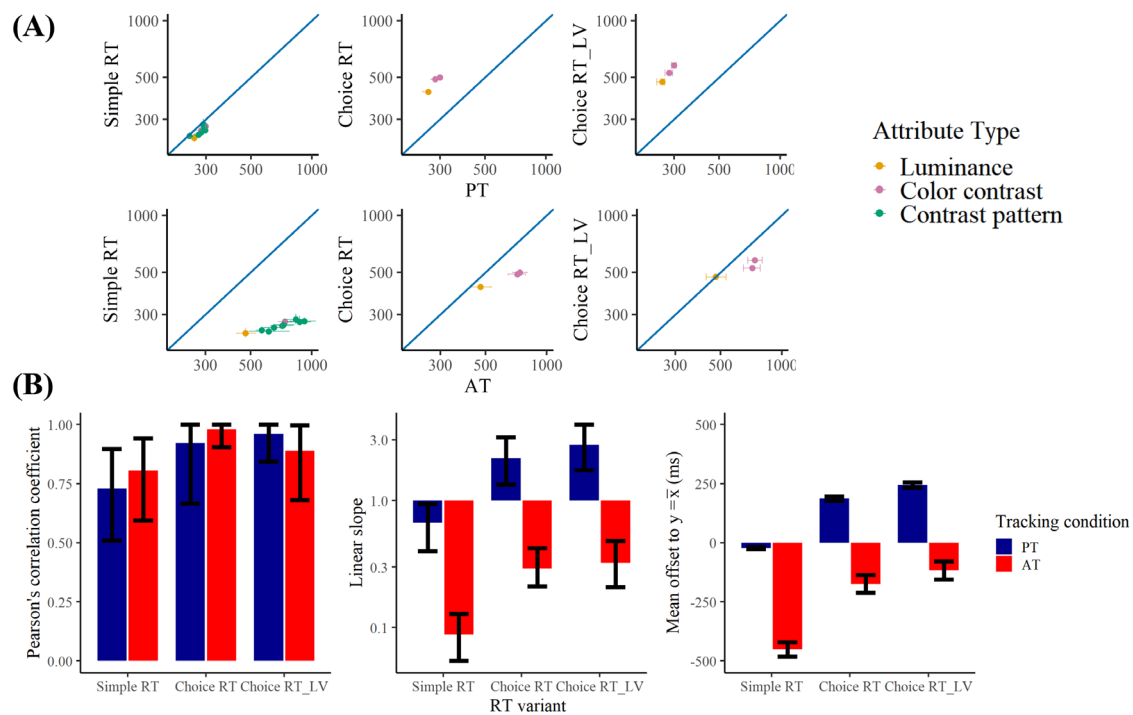


Figure 6-10. Experiment 4-3 result.

(A) The graph displays the peak latencies of PT and AT in relation to various RT values. The top row illustrates RT versus PT, while the bottom row shows RT versus AT. Moving from left to right, the columns represent simple reaction time, choice reaction time, and low-value choice reaction time, respectively. A blue diagonal line represents an identity line with a slope of one and a zero intercept. Different attribute types are depicted by dots of varying colors. Each dot is accompanied by error bars, which indicate the 95% confidence interval between the 2.5 and 97.5 quantiles. (B) The similarity measures between AT/PT and RTs. The graph from left to right shows the Pearson correlation, linear slope, and mean offset, respectively. The error bars associated with each bar indicate the 95% confidence interval between the 2.5 quantile and 97.5 quantile.

6.5. General Discussion

The study directly compared the temporal dynamics of position and attribute tracking (PT and AT) for luminance, color contrast, and luminance contrast. The findings demonstrated that continuously tracking the attribute value of the stimuli (AT) is slower than tracking the location of the stimulus (PT), but the sharing of the encoding process between PT and AT is also suggested by a positive correlation in the effects of stimulus attribute on the peak latency. The subsequent experiment that altered the Stimulus-Response mapping and local variation magnitude of PT successfully aligned the latencies of PT close to those of AT. The result demonstrated that the disparity in the underlying

mechanism between PT and AT comes from a difference in the S-R mapping, in addition to a difference in the effective stimulus strength. The reaction time measurements provide additional evidence. The simple reaction time that primarily evaluates the basic speed to detect a stimulus change is closely associated with the peak latency of PT, while the choice reaction time that includes additional processes is closely associated with the peak latency of AT.

6.5.1. Common attribute encoder for PT and AT

In the PT task, participants were required to monitor a dynamic change in the target's position. In the AT task, in contrast, participants were required to monitor a dynamic change in the attribute value of the target stimulus. In spite of the difference in the task objective, our findings revealed clear correlations in the effects of stimulus attribute on the peak latency between AT and PT or its variants (PT_LV, PT_InvLV) except for PT_Inv. This suggests that PT and AT shares a mechanism that detects attribute changes (i.e., the attribute detector). This relationship may stem from the partial overlap in task objectives between PT and AT: both require the visual system to encode the local variation of the attribute value.

As shown by the slope of the scatter plots in Figure 6-6, the relative magnitude of the effects of stimulus attribute on the peak latency was much smaller for the original PT than for AT. The slope became nearly zero when a constant offset and noise were added to the latency by changing the task into the antireaction task (PT_Inv). However, a reduction of the stimulus intensity of PT (PT_LV, PT_InvLV) increased the magnitudes of the effect of stimulus attribute up to a level comparable to that of AT. This finding indicates that the low gain of the original PT does not reflect not the involvement of a distinct attribute encoder in PT, but a saturation of the effects of stimulus intensity on the latency shortening for PT (Barnett et al., 2024). That is, the effective stimulus intensity of the original PT was so high that the effect of stimulus difference was hindered by compressive nonlinearity (Barnett et al., 2024)

6.5.2. Stimulus-Response mapping

We propose that one factor producing the slowness of AT, relative to PT, is that the AT task includes a voluntary stimulus-response (S-R) mapping. In contrast, PT can be assisted by a mechanism of involuntary and rapid hand movement following a visual

target (Pisella et al., 2000). Measurements of reaction time yield comparable findings. While AT is more closely associated with choice reaction time, PT demonstrates a stronger connection to detection reaction time. The primary difference between these two types of reaction times is the requirement for a content-“awareness” mechanism. This suggests that the PT task, along with the simple detection task, is not significantly influenced by the specific content of the stimuli, whereas AT, along with the choice reaction task, needs to process the stimulus content to choose a proper response. Combining the results from both tracking and reaction time tasks indicates that AT differs from PT with respect to the inclusion of content-aware processing in the stimulus-response mapping process.

In accordance with the similar temporal latency between AT and antireaction PTs, a similar voluntary motor mapping system may be included; however, the underlying mechanism between PT_Inv and AT can be different. For example, when performing the antireaction task in PT_Inv or PT_InvLV, the system may be required to inhibit the original S-R mapping and then voluntarily establish a new one, whereas in AT, it could be simply related to voluntary S-R mapping. In this regard, however, recent studies have shown that voluntary inhibition of an involuntary motor command involves neural regions common to voluntary motor commands (Ghosh et al., 2014; Kenner et al., 2010), suggesting the difficulty in distinguishing between inhibition of the original S-R mapping and the voluntary construction of a new S-R mapping. While we attributed the increased latency primarily to the voluntary and cognitive demands of inverse mapping, prior research indicates that different misalignment angles may have varying effects on sensorimotor integration (Ellis et al., 2013; Ellis & Adelstein, 2014). Therefore, an important next step would be to systematically investigate how sensorimotor misalignment at different angles interacts with cognitive load in continuous tracking tasks, potentially providing deeper insights into the underlying mechanisms of PT and AT differences.

Moreover, we examined whether participants could improve their performance in these more demanding tracking tasks through practice. However, we found no evidence of a learning effect across the 30 trials in each condition, as tracking latency did not significantly decrease over time. This may reflect both the short timescale of the task and the lack of informative feedback, particularly in the inverted mapping condition, where participants did not receive explicit cues to support learning a reversed mapping

strategy. One more limitation is that, since the current study only used hand movement to collect the response, it remains unknown whether we can still observe the PT-AT consistency when we use other motor responses, such as eye movement.

6.5.3. Continuous tracking performance

This study compared the peak latencies for two types of continuous tracking and the reaction times for similar stimuli. Their mutual relationships can be approximated as follows:

$$AT \approx 5.33 PT + 400 \text{ (ms)}$$

$$RT_{\text{simple reaction}} \approx 1.22 PT - 22.5 \text{ (ms)}$$

$$RT_{\text{choice reaction}} \approx 0.4 AT - 117 \text{ (ms)}$$

Despite being obtained under a specific condition, these functions will be useful for comparison the results obtained with different psychophysical tasks.

Peak latency is not the sole metric for analyzing the CCG. Other indices such as half-height bandwidth or peak amplitude may also be informative; however, we observe a high level of consistency among these measurements. The Pearson's correlation coefficient is 0.98 between peak latency and bandwidth, -0.83 between peak latency and peak amplitude, and -0.88 between bandwidth and peak amplitude. This result aligns with Bonnen et al., (2015), which suggested that these three indices are highly consistent, with their measurements showing a linear correlation of approximately 0.99. Considering these findings, we posit that these three indices represent similar aspects of the system's temporal characteristics, at least within the context of our measurements.

Focusing solely on peak latency data might lead to the misconception that CCG is simply another rapid variant of the reaction time paradigm. Subsequent research could explore the underlying mechanisms responsible for shaping the CCG function. To achieve this, a comprehensive dynamic model integrating processes from perception to action would be necessary.

One methodological point to note is that our trajectory generation process involved rejecting approximately 70–80% of the randomly generated sequences, based on criteria ensuring sufficient drifting across the space (spatial or attribute value). While this

filtering step deviated from the original generative model, it did not affect the validity of our CCG analysis, as the retained trajectories still exhibited near-delta autocorrelation functions. However, future modeling efforts that directly rely on the assumed stimulus distribution, for example, those simulating or analyzing behavioral trajectories—should consider the empirical distribution of the final stimuli rather than the idealized generative process.

6.5.4. Potential Model for tracking task

While the present study focused on data-driven comparisons between position tracking (PT) and attribute tracking (AT), we recognize that a formal computational model could provide a more unified account of the mechanisms underlying the observed latency differences. Such a model would need to integrate multiple stages (Brueggemann, 2007)—from perceptual encoding to motor execution—possibly with distinct latency parameters for position and attribute signal, which will link to the traditional psychophysics measured impulse response (Burr & Morrone, 1993; Shinomori & Werner, 2003; Watson, 1986), as well as a cost function associated with control effort and mapping complexity (Bonnen et al., 2015; Straub et al., 2023; Straub & Rothkopf, 2022).

Existing models of continuous tracking, such as the Kalman filter framework proposed by Bonnen et al., (2015), and the Bayesian optimal control model by Straub & Rothkopf, (2022), primarily explain the shape of the CCG based on perceptual and motor uncertainty in stimulus position. These uncertainties influence the dynamics of motor output through time-dependent transition matrices. However, Bonnen et al., (2015) concurrently reduced the peak intensity with an increase in the Gaussian spread, also Straub & Rothkopf, (2022) use their data for modelling. The results of our experiments, as well as those by Barnett et al., (2024) suggest that an increase in temporal integration time accompanied by an effective stimulus contrast reduction may also contribute to the results reported by Bonnen et al., (2015). Future studies may wish to systematically investigate how variations in stimulus magnitude contribute to temporal uncertainty in continuous tracking paradigms (both for position and attribute tracking).

As an initial attempt to address this issue, we introduced a simple linear model that decomposes the peak latency into components attributed to stimulus encoding and S—

R mapping demands. This model, estimated using least squares on group-averaged data from 6 participants, successfully captured the relative contributions of attribute type, magnitude, and motor mapping, achieving a high degree of fit ($R^2 = 0.97$). While this approach provides a compact explanation for peak latency differences across conditions, it does not yet account for the full shape of the CCG—particularly its bandwidth, slope asymmetry, or delayed onset—features that likely reflect richer temporal dynamics within perceptual and motor systems.

A comprehensive future model should explicitly dissociate the contributions of spatial and temporal uncertainty, and possibly incorporate individual differences in cost function strategies, as suggested by Straub & Rothkopf, (2022). We see our current findings as laying the empirical groundwork, along with a rich dataset, to support and constrain such future modeling efforts.

6.6. Conclusion

This research introduces continuous psychophysics paradigm to examine various attributes' temporal dynamics, aiming to investigate the different processes between position tracking (PT) and attribute tracking (AT). Findings reveal that AT is significantly slower than PT to complete the task, yet the latency difference among different attributes shows a strong correlation, due to the use of a common attribute encoder. Lower stimulus variation magnitude and voluntary stimuli-response mapping can explain the difference of AT from PT. Moreover, the measurements by this new paradigm correspond well with traditional reaction time measurements.

Chapter 7. Temporal dynamics of temporal correlation-based visual segmentation

Following the methodological groundwork established in Chapter 6, this chapter advances the investigation by focusing on the temporal dynamics underlying perceptual segmentation. Using the tracking-based behavioral data, we implement a computational modeling approach to examine how the system responds over time under different processing demands. Specifically, this chapter focuses on modeling the differing temporal dynamics between conditions that require temporal correlation computation and those that do not, with the aim of uncovering how temporal information is integrated within the system across distinct task contexts.

7.1. Introduction

Temporal correlation between local visual features has long been recognized as an important cue for perceptual grouping and segmentation. When nearby regions exhibit synchronized temporal changes, which reflect high temporal correlation, observers are more likely to perceive them as a unified area. In contrast, asynchronous changes tend to promote perceptual separation (Chapter 3, Chapter 5, Guttman et al., 2005; Lee & Blake, 1999). This phenomenon has been demonstrated across a range of visual attributes, including luminance (Usher & Donnelly, 1998), contrast (Alais et al., 1998; Guttman et al., 2005), color (Leonards & Singer, 1998), orientation (Leonards et al., 1996), motion direction (Kandil & Fahle, 2004; S.-H. Lee & Blake, 1999). Moreover, this effect occurs regardless of whether the stimulus is defined by a single attribute or by multiple attributes presented simultaneously (Chen et al., 2024a; Guttman et al., 2005). These findings suggest the presence of a general mechanism that is sensitive to temporal correlation and contributes to the organization of visual input into coherent perceptual units.

Although computing temporal correlation requires comparing signals across more than one point in time, this does not necessarily imply that the visual system accumulates correlation values as part of its decision-making process. The structural requirement of comparing multiple moments does not in itself indicate that temporal

correlation is explicitly encoded or integrated by the system. An alternative possibility is that the system detects momentary changes in temporal contrast and initiates a response when the contrast becomes strong enough to exceed an internal threshold. In this scenario, the system would respond to immediate signal fluctuations without computing or accumulating correlation over time (Adelson & Farid, 1999; Farid, 2002).

To test whether the visual system truly accumulates temporal correlation over time to support segmentation, we conducted Experiment 5-1. In this experiment, we examined how reaction time varies across different levels of temporal correlation. Specifically, we measured the time required for observers to detect a segmentation boundary when the correlation strength between visual elements was manipulated. To eliminate confounding effects from local temporal contrast, we used stimuli developed in our previous study (Chen et al., 2025), that were designed to have a flat temporal power spectrum. These stimuli contained equal energy across all temporal frequencies and did not contain salient temporal boundaries that could serve as simple cues for segmentation. If the system relies on accumulating temporal correlation, stronger correlation should lead to faster segmentation responses. In contrast, if segmentation is based on local contrast cues, then reaction time should remain relatively constant regardless of the correlation strength.

While reaction time measures the minimum amount of time required to initiate a response, it does not reveal how much information is integrated before the response is triggered. In other words, reaction time does not indicate the length of the temporal integration window or the extent of evidence that contributes to the final decision. To address this limitation, we employed a continuous psychophysics approach (Bonnen et al., 2015), in which participants were asked to continuously track the position of a target stimulus. This method has been shown to be sensitive to the encoding dynamics of visual attributes such as contrast (Bonnen et al., 2015), depth (Bonnen et al., 2017; Burge & Cormack, 2024), and color (Barnett et al., 2025). In our previous studies, we found that the temporal profile derived from such tracking tasks can be decomposed into two components: perceptual encoding and motor mapping. The duration of the encoding component remains consistent across tasks involving position tracking and those involving feature tracking (Chen et al., 2024b). Since the tracking task produces a continuous stream of behavioral output in response to ongoing visual input, it allows us to estimate the full time course of processing from perception to action. Using this

method, we aim to infer the temporal integration window associated with the computation of temporal correlation.

Based on this framework, Experiment 5-2 was designed to further isolate the contribution of temporal correlation. We created three conditions that varied in task complexity. The first condition involved detecting the target's location, which required minimal interpretation of temporal dynamics. The second condition required discriminating temporal contrast between the target and background. The third and most complex condition required distinguishing differences in temporal correlation structure between the target and background. Using data from these tasks, we developed a computational model to estimate the temporal dynamics of each process involved. This allowed us to isolate the integration window that specifically reflects sensitivity to temporal correlation. Our goal was to clarify whether correlation-based segmentation relies on a distinct accumulation process and to quantify the temporal span over which relevant information is integrated.

Taken together, the present study aims to investigate the temporal integration window that supports segmentation based on temporal correlation. We approach this question through two complementary experiments. In Experiment 5-1, we examined whether different levels of temporal correlation influence reaction time, which reflects the minimum latency required to trigger a segmentation response. In Experiment 5-2, we measured the full temporal dynamics of perceptual processing using a continuous tracking paradigm. By applying a computational model to the behavioral data, we estimated the temporal integration window associated with correlation-based segmentation. To assess whether this integration process is consistent across different visual features, we tested three types of stimulus elements: luminance-defined blobs, red–green color blobs, and blue–yellow color blobs. This allowed us to compare the temporal integration profiles across attributes that differ in both low-level encoding and potential neural substrates. By examining the width and structure of the integration window across these feature domains, our findings provide new evidence regarding the degree of feature invariance in correlation-sensitive segmentation mechanisms, offering insights into a long-standing question about the generality and specificity of temporal grouping in human vision.

7.2. Experiment 5-1. The reaction time paradigm

7.2.1. Methodology

7.2.1.1. Apparatus

The stimuli were displayed on a VIEWPixx /3D LCD monitor (VPixx Technologies, Saint-Bruno-de-Montarville, Canada) with a resolution of $1,920 \times 1,080$ pixels and a refresh rate of 30 Hz. The lowest, mean, and highest display luminance ranged from 1.8 cd/m^2 (lowest) to 96.7 cd/m^2 (highest), with a mean luminance of 48.4 cd/m^2 . The linear output of luminance was calibrated for each channel using an i1Pro chromometer (VPixx Technologies). Each pixel subtended 1.3 arcmin at a viewing distance of 70 cm. Participants were seated in a dark room with a chinrest to stabilize head position. The experiments were programmed using the Psychophysics Toolbox (Brainard, 1997) in MATLAB release 2022a (MathWorks, Natick, MA, USA). The study protocol adhered to the ethical standards of the Declaration of Helsinki, except for preregistration, and was approved by the Ethics Committee of Kyoto University (KUIS-EAR-2020-003).

7.2.1.2. Participants

The experiment included two of the authors and four naïve participants (six males; mean age: 25 years), all of whom reported normal or corrected-to-normal vision. Participants were informed of the study's nature and provided written informed consent before participation. Monetary compensation was provided upon completion. Prior to the main experiment, a minimum motion method (Anstis & Cavanagh, 1983; Cavanagh et al., 1987) was used to measure the equiluminant ratio (in terms of the system intensity value) between red (R) versus green (G) and blue (B) versus yellow (Y) for each participant before the main experiment. During the procedure, participants saw a Gabor drifting rightward. The sinusoidal waves of R and G or B and Y are 180° out-of-phase. The participants were required to adjust the ratio between the channels to minimize their sense of motion. The mean R: G ratio value was 0.58, and the mean B: Y ratio was 0.20.

7.2.1.3. Stimuli

We utilized the feature tensor developed by Chen et al. (2025), in which each tensor consists of an 8×16 spatial grid spanning 30 temporal frames. The values within the tensor range from -1 to 1 . Each tensor can be divided into left and right regions based on the degree of temporal correlation specified within each half. We denote the tensor as $F(i, j, t; \phi)$, where (i, j) refers to the spatial position, t indicates the time frame, and ϕ specifies the magnitude of temporal correlation. For each of the 11 correlation levels ranging from 0 to 1, a total of 30 tensors were generated. This layout allows the stimulus to be divided into two 8×8 regions based on the temporal correlation assigned to each half.

Each stimulus was constructed by randomly selecting one tensor F from the corresponding set. The stimulus took the form of an 8×16 dynamic lattice, presented for 30 frames, corresponding to a 1-second presentation. Each element in the lattice subtended 0.5 degrees of visual angle, resulting in an overall stimulus size of 4 degrees in height and 8 degrees in width. The values in F were mapped to individual lattice elements, with each element corresponding to one of three types of Gaussian-shaped patterns: a luminance blob, a red–green (Color-R/G) blob, or a blue–yellow (Color-B/Y) blob. A representative appearance of the stimulus can be found in Figure 7-1 (A). The stimulus profile of a single element is described as follows:

$$\begin{bmatrix} R(x, y, t) \\ G(x, y, t) \\ B(x, y, t) \end{bmatrix} = L_{mean} \cdot \left(1 + \delta \cdot F(i, j, t) \cdot \begin{bmatrix} C_R \\ C_G \\ C_B \end{bmatrix} \right) \cdot e^{-\frac{x^2+y^2}{\sigma_T^2}} \quad (7-1)$$

x and y denote the spatial coordinates within the element, centered at zero. σ_T is the Gaussian window of the target, which was set to 0.1 degrees. δ denotes the polarity of the signal and was randomly assigned a value of either -1 or 1 for each element, inde-

pendently sampled. The color contrast vector $\begin{bmatrix} C_R \\ C_G \\ C_B \end{bmatrix}$ specifies the chromatic weighting of

the pattern. For luminance blobs, it was defined as $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$; For red–green blobs, it was

defined as $\begin{bmatrix} 1 \\ -L_{Iso.RG} \\ 0 \end{bmatrix}$, and for blue–yellow blobs, it was defined as $\begin{bmatrix} L_{Iso.BY} \\ L_{Iso.BY} \\ -1 \end{bmatrix}$. $L_{Iso.RG}$

is the iso-luminant index between red and green while $L_{ISO.BY}$ is that between blue and yellow.

Each trial began with a dynamic white noise sequence that preceded the presentation of the correlation-defined stimulus described above, a timeline can be found at Figure 7-1 (B). This white noise segment was independently generated for each trial and followed the same spatial and temporal structure as the main stimulus. Specifically, the noise tensor $F_N(i, j, t)$ was sampled from a uniform distribution with values drawn from $[-1, 1]$ over an 8×16 spatial grid and a temporally varying duration T , which was randomly selected from the discrete uniform distribution $T \sim Unif\{15, 16, \dots, 36 \text{ frames}\}$. This corresponds to a variable duration of 500 to 1200 milliseconds. The white noise segment was rendered using the same stimulus generation procedure as described in Equation (7-1), where the values in F_N were used to modulate either luminance or color blobs depending on the element type. As a result, the white noise sequence was perceptually indistinguishable in format from the subsequent correlation-defined stimulus but contained no structured temporal correlation. This design ensured that the onset of signal information was temporally unpredictable and prevented participants from anticipating the appearance of the relevant segmentation cues.

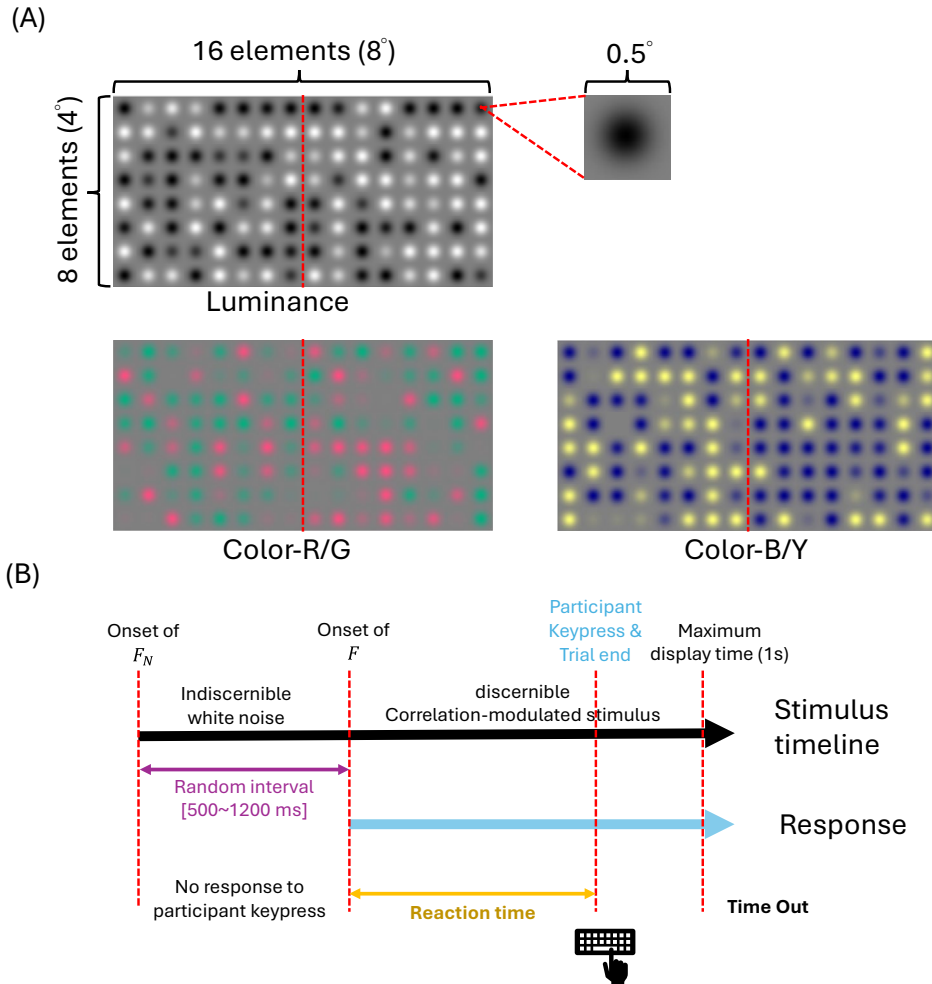


Figure 7-1. Stimulus structure and trial timeline in the reaction time task.

(A) Example stimuli used in the reaction time experiment. Each stimulus consisted of an 8×16 element lattice, where each element subtended 0.5° and was defined as either a luminance Gaussian blob (top), a red–green color contrast blob (bottom left), or a blue–yellow color contrast blob (bottom right). A vertical red line indicates the midline division between the left and right 8×8 regions show the potential segmentation boundary. Temporal correlation was manipulated within each half of the stimulus. (B) Schematic of the trial structure. Each trial began with a white noise stimulus F_N , generated from uncorrelated spatiotemporal signals and presented for a random duration between 500 and 1200 milliseconds. This was followed by the onset of a correlation-modulated stimulus F , which participants viewed until they detected a segmentation cue. Reaction time was measured from the onset of F to the participant's keypress. Trials ended upon response or after a maximum display time of 1 second. Premature responses during the white noise period were not accepted.

7.2.1.4. Design and procedure

We measured the reaction time (RT) for detecting segmentation boundaries across 11 levels of within-area temporal correlation, ranging from 0 to 1 in equal steps. The within-area correlation referred to the temporal consistency of elements within each half of the 8×16 stimulus lattice. In each trial, both the left and right halves shared the same correlation level, while the correlation between the two halves was fixed at zero. For each correlation level, 90 trials were collected. Each trial started with a white

noise stimulus that lasted for a random duration between 500 and 1200 milliseconds. The noise was temporally unstructured and visually indistinguishable in terms of segmentation. Following the noise period, a correlation-modulated stimulus appeared, which could potentially be segmented into two regions based on temporal dynamics. Participants were asked to press the spacebar as soon as they perceived a segmentation boundary. The stimulus disappeared either upon keypress or after 1 second. If participants did not respond within this 1-second window, the trial was marked as a timeout and recorded as 9999. These timeout trials were excluded from the later calculation the average RT. A timeline of one trial demonstration can be referred to Figure 7-1 (B). Each participant completed three types of stimuli: luminance, red–green color contrast, and blue–yellow color contrast. In total, participants completed $3 \text{ types} \times 11 \text{ correlation levels} \times 90 \text{ trials} = 2970 \text{ trials}$. The entire experiment took approximately two and a half hours to complete. The condition sequence was randomized across trials.

7.2.2. Result

We observed a consistent pattern across all three stimulus types. As the temporal correlation increased, participants were more likely to report seeing a clear segmentation between the two regions. This trend was reflected in the proportion of selection (%selection), which increased monotonically with correlation strength (Figure 7-2, black lines). In terms of reaction time, a noticeable decrease in RT was observed when the correlation exceeded a moderate threshold, typically between 0.4 and 0.5. Below this level, the average RT showed a slight upward trend. However, this increase should be interpreted with caution, as the RT distributions under low correlation conditions (shown as light blue density plots in Figure 7-2) appeared relatively flat and uniform, lacking a distinct central tendency. This suggests that the apparent increase in mean RT under low-correlation conditions does not reflect a meaningful or reliable pattern. For luminance stimuli, the shortest RT was 532.76 ms at correlation 1.0, and the longest was 683.89 ms at correlation 0.4. For red–green stimuli, the shortest RT was 615.26 ms at correlation 1.0, and the longest was 751.02 ms at correlation 0.5. For blue–yellow stimuli, the shortest RT was 532.76 ms at correlation 1.0, and the longest was 710.40 ms at correlation 0.4. The experimental results suggest that the strength of temporal correlation not only influences the likelihood of perceiving segmentation, but also directly modulates the amount of time needed for segmentation to occur.

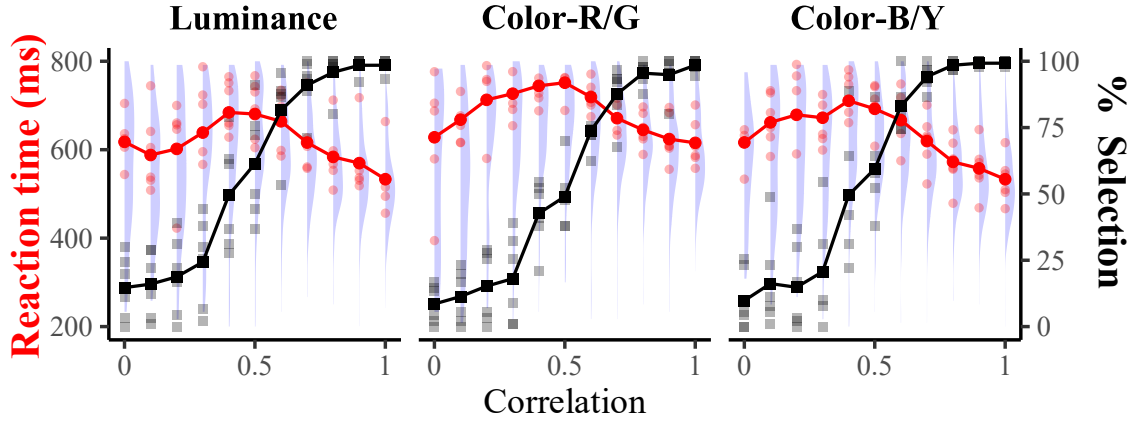


Figure 7-2. Reaction time and selection rate across levels of temporal correlation.

Reaction time (RT) and response rate are plotted as a function of within-area temporal correlation for three stimulus types: luminance, red–green (Color-R/G), and blue–yellow (Color-B/Y). The red line with circular markers shows the average RT across six participants. The black line with square markers shows the average proportion of valid responses (keypresses within 1 second). Light red dots represent individual RT data for each participant, and gray squares represent individual response rates. The light blue violin plots indicate the distribution of RTs concatenated across participants at each correlation level.

7.3. Experiment 5-2. The continuous tracking paradigm

7.3.1. Methodology

7.3.1.1. Stimuli

Using the stimulus generation protocol developed in our previous work (Chen et al., 2025), we created a new set of feature tensors with a spatial resolution of 48×48 elements and a temporal length of 600 frames. A total of 30 variants were generated. Each tensor can be divided into two spatial halves, with the left and right regions consisting of 48×24 elements each (1152 elements per half). Within each half, the temporal correlation across elements was set to 1.0, while the correlation between the two halves was set to 0. We refer to these two sub-tensors as F_L (left half) and F_R , respectively. In the subsequent stimulus generation process, elements sampled from F_L were used to define the target region, while those from F_R served as the background. Because the target moves across the display during the task, special care was taken to eliminate unintended temporal contrast cues that might arise from transitions between target and background signals. To achieve this, we introduced an additional constraint during tensor generation: we ensured that the temporal value differences between consecutive frames were matched across all spatial locations, regardless of whether the change

occurred within a single region or across regions. This control was implemented to avoid introducing local temporal transients that could serve as potential segmentation cues.

We then generated the feature tensor used for the tracking task, denoted as F_{Track} , by superposing two surfaces: a target surface S_T , sampled from F_L , and a background S_B sampled from F_R . The combination was weighted by a circular mask with a diameter of 12 elements (corresponding to 6 degrees of visual angle), centered at a position that drifted horizontally over time. The formal expression of the stimulus tensor is given as:

$$F_{Track}(h, v, t) = A(P(t)) \cdot S_T(h, v, t) + (1 - A(P(t))) \cdot S_B(h, v, t) \quad (7-2)$$

Here, t represents time, spanning 600 frames at 60 Hz, resulting in a total stimulus duration of 10 seconds. The mask center position $P(t)$ follows a Brownian motion trajectory defined as:

$$P(t) = \left[\sum_{t'=0}^t V(t') \right] + 42 \quad (7-3)$$

$$V(t') \sim \mathcal{N}(0, \sigma_P^2) \quad (7-4)$$

The output of $P(t)$ is rounded to the nearest integer to ensure that the target shifts by whole-element units. The constant 42 shifts the starting location from zero to the horizontal center of the display (the 42nd column). The standard deviation of the motion, σ_P was set to 0.69 elements per frame, which corresponds to approximately 20.75 arcminutes per frame. Several constraints were imposed on the generated trajectory to ensure visibility and spatial confinement. First, $P(t)$ was required to stay within the range [6, 77], ensuring that the target remained fully within the display bounds. This limited the maximum horizontal movement range to 71 elements (35.5 degrees). Second, the total movement over the 10-second period was required to exceed 20 elements (10 degrees) to guarantee that the target displacement was clearly observable. The overall rejection rate of trajectories failing to meet these constraints was around 66%.

The circular mask function was defined as follows:

$$A(h, v, t, P(t)) = \begin{cases} 1, & D(h, v, P(t)) \leq 6 \\ 0, & \text{otherwise} \end{cases} \quad (7-5)$$

$$D(h, v, P(t)) = \sqrt{(h - P(t))^2 + (v - 23)^2}$$

Under this masking configuration, 112 elements were located within the circular region, and 3706 elements were outside.

The target and background surfaces S_T and S_B were generated by randomly sampling spatial positions from F_L and F_R , respectively, and assigning their entire temporal sequences to corresponding positions in the surface. Specifically:

$$\begin{aligned} \{S_T(h, v) &\leftarrow F_L(i, j, 1:600) \\ \{S_B(h, v) &\leftarrow F_R(i, j, 1:600) \\ \forall i, k &\sim \text{Unif}(\{1, 2, \dots, 83\}) \\ \forall j, l &\sim \text{Unif}(\{1, 2, \dots, 23\}) \end{aligned} \quad (7-6)$$

Finally, the feature tensor was transformed into displayable stimuli using the same rendering protocol described in Experiment 1.

7.3.1.2. Design and procedure

Three tracking conditions were designed based on the stimulus configuration, and we show the demo appearance of this three condition in Figure 7-3.

- **Dynamic background:** This condition used the full configuration defined in Equation(7-6), where both the target and the background varied over time following a quasi-white noise pattern. In this setting, no temporal contrast is available between the target and the background, making the computation of temporal correlation necessary for successful tracking.
- **Fixate background:** In this condition, the background surface S_B as held constant throughout the trial. Each spatial location was assigned a fixed value drawn from a uniform distribution between -1 and 1 , and the same values were presented across all 600 frames. Here, temporal correlation computation is not required, but detecting the target relies on the presence of temporal contrast between the dynamic target and the static background.

- **Plain background:** In this condition, the background was set to a uniform gray level across space and time. Since both spatial and temporal variation in the background were removed, only spatial contrast between the target and background were removed, only spatial contrast between the target and background was available as a cue for tracking. Temporal contrast computation was not necessary in this condition.

Each participant completed 30 trials for each combination of stimulus type and tracking condition. As in Experiment 5-1, three stimulus types were tested: luminance, red–green (Color-R/G), and blue–yellow (Color-B/Y). This resulted in 3 stimulus types $\times$ 3 tracking conditions $\times$ 30 trials = 270 trials in total. Each trial lasted 10 seconds. The full experiment required approximately 2.5 hours to complete, including short breaks between blocks. The order of conditions was randomized across participants.

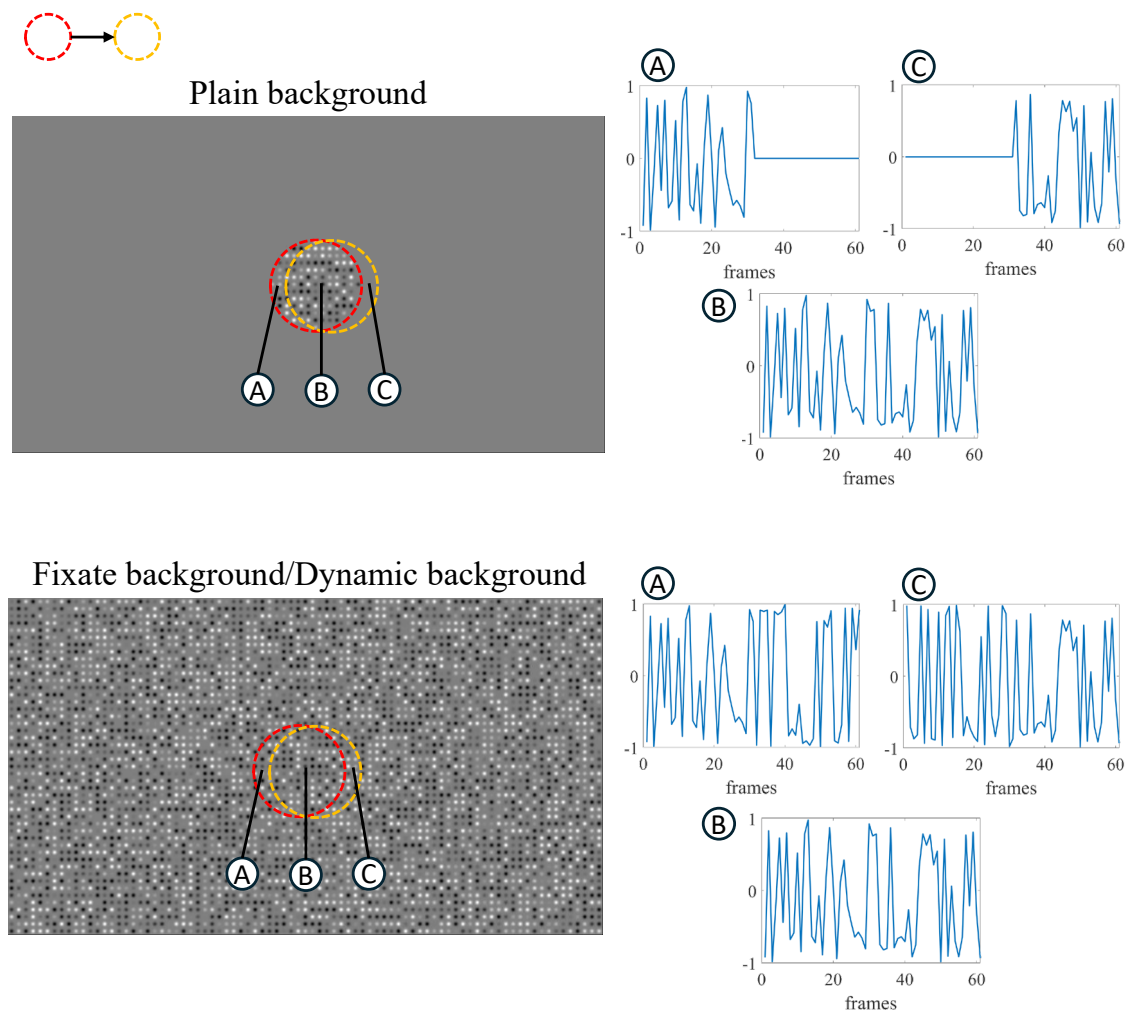


Figure 7-3. Demo spatial and temporal waveform of three conditions.

The figure illustrates the spatiotemporal structure of stimuli across different tracking conditions. The upper panel shows the plain background condition, while the lower panel applies to both fixated and dynamic background conditions. The two solid circles represent the target area at two different time points, with the red circle indicating an earlier position and the yellow circle indicating a later one. Note

that this shift is schematic and does not correspond to actual stimulus motion. Panels A, B, and C show the 60-frame temporal sequences of three representative spatial locations. Panel A corresponds to a position that transitions from target to background, Panel B remains within the target region, and Panel C transitions from background to target. In the plain background condition (upper row), the transition points produce clear and abrupt temporal changes in Panels A and C. In contrast, in the dynamic background condition (lower row), no clear temporal transients are detectable following the transition, as the temporal waveforms remain noisy and statistically indistinguishable. The fixated background condition (not shown separately) would appear similar to the plain background condition, but with random temporal offsets at each location.

7.3.1.3. Analysis

We use the normalized cross-correlation with the time-shifting velocity between pixelized target trajectory $P'(t)$, which timing the pixel size of the original $P(t)$ to transform from element-based into pixel-based, and the velocity of participants' response trajectory $R(t)$ also in pixel-based for per single trial and then took the average over 30 trials to compute the cross-correlogram (CCG):

$$CCG(\tau) = \frac{1}{30} \sum_{k=1}^{30} G_{R \cdot P',k}(\tau) \quad (7-7)$$

$$G_{R \cdot P,k}(\tau) = \frac{Cov_{R \cdot P',k}(\tau)}{\sqrt{Cov_{R \cdot R',k}(\tau) Cov_{P' \cdot P',k}(\tau)}}, \quad (7-8)$$

$$\tau \in [-n + 1, -n + 2, \dots, n - 1]$$

$$Cov_{R \cdot P,k}(\tau) = \begin{cases} \sum_{i=1}^{n-\tau} \Delta R_{i+\tau,k} \Delta P'_{i,k}, & \tau \geq 0 \\ \sum_{i=1}^{n+\tau} \Delta R_{i,k} \Delta P'_{i-\tau,k}, & \tau < 0 \end{cases} \quad (7-9)$$

Here, the differentiation operator $\Delta(X) = X_i - X_{i-1}$ captures the instantaneous velocity of the response and the target. The first 2 seconds of the data were removed to exclude unstable movements. For each condition of each participant, we had 30 $G_{R \cdot P',k}(\tau)$ and one CCG. We used the bootstrapping to estimate the standard error of CCG. Specifically, at each iteration of the bootstrapping, we resampled 30 $G_{R \cdot P',k}(\tau)$ with replacement, and then calculated the CCG. The bootstrapping process was repeated for 2500 times of iteration (i.e., $CCG_1, CCG_2, \dots, CCG_{2500}$). The normalization operation follows the Mulligan (1997) to ensure the correct estimation of the coefficients of correlation for each frame.

7.3.2. Result

We present the cross-correlation results for all stimulus conditions and participants in Figure 7-4. CCG result across participants on different attribute and tracking conditions.. The dynamic background condition (red line) consistently exhibited slower and more difficult tracking performance compared to the other two conditions. This was reflected in a lower peak amplitude, longer peak latency, and broader temporal bandwidth of the cross-correlation functions. This pattern was observed across all three visual attributes and was consistent across individual participants. Additionally, there was no apparent difference in peak latency between the fixated background and no-background conditions. A small latency advantage of approximately 20 ms was observed in the no-background condition, suggesting slightly faster responses in the absence of background modulation (refer to Table 7-1. Peak latency (in milliseconds) across different attribute types and tracking conditions.). Minor differences were also observed across visual attributes, although the pattern was not consistent across background conditions. In the no-background and fixated background conditions, tracking based on luminance and blue–yellow attributes tended to produce slightly earlier peaks than red–green. In contrast, under the dynamic background condition, red–green tracking responses exhibited marginally shorter latencies than the other two. These results suggest that while background dynamics are the primary factor influencing tracking latency, subtle interactions with feature type may also modulate temporal integration in a condition-dependent manner.

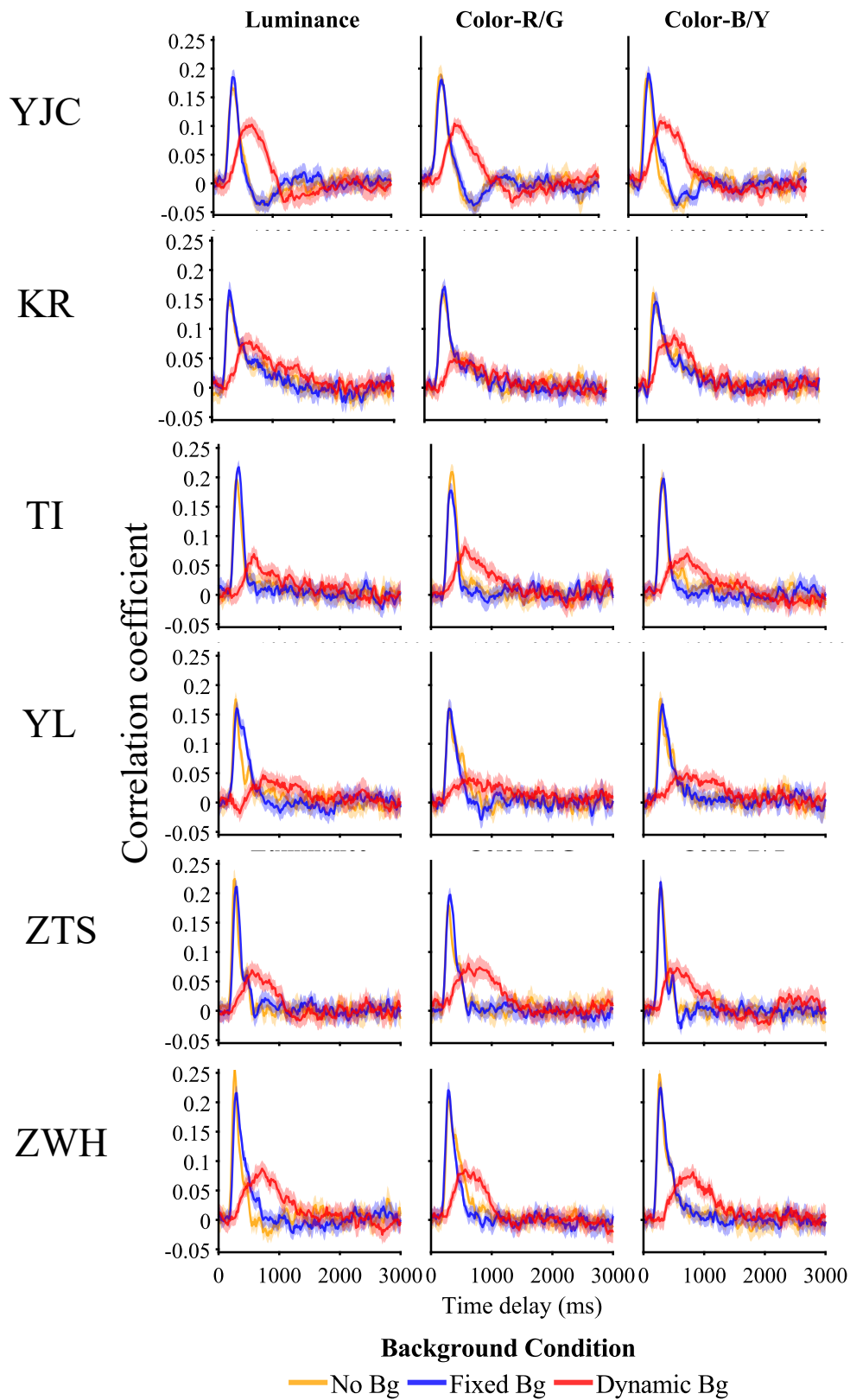


Figure 7-4. CCG result across participants on different attribute and tracking conditions. Cross-correlation functions (CCGs) between the target position and the participant's tracking response for six individual participants (rows) across three visual attributes: luminance, red–green (Color-R/G), and blue–yellow (Color-B/Y) (columns). Each plot shows the CCG curves for three background conditions: plain background (yellow), fixed background (blue), and dynamic background (red). The x-axis

represents the time delay (in milliseconds), and the y-axis shows the correlation coefficient. Shaded areas indicate 95% confidence intervals computed by bootstrapping the trial-level data (2.5 to 97.5 percentiles).

Table 7-1. Peak latency (in milliseconds) across different attribute types and tracking conditions.

Peak latency (SE)	Plain background	Fixate background	Dynamic background
Luminance	286.11 (7.26)	308.33 (7.61)	641.67 (27.15)
Color-RG	311.11 (9.35)	319.44 (9.14)	569.44 (19.80)
Color-BY	283.33 (4.46)	311.11 (8.49)	661.11 (39.80)

Note. Values indicate the average peak latency across participants for each condition, measured in milliseconds. Values in parentheses represent the standard error (SE) of the mean.

7.4. Computational model: integrating RT and tracking data

Across the previous two experiments, we demonstrated that the visual system computes and accumulates temporal correlation signals to support segmentation (Experiment 5-1), and that such computations introduce a measurable delay in behavioral responses compared to tasks that do not require temporal correlation processing (Experiment 5-2). These findings suggest that an additional mechanism, sensitive to temporal correlation, is engaged when the system must distinguish a target embedded within a dynamic background. To address our original question regarding the amount of temporal information involved in this correlation-based process, we developed a computational model that simultaneously accounts for both the reaction time and tracking data. This model provides a quantitative index representing the temporal integration window, which characterizes the duration over which correlation-related evidence is accumulated in the visual system.

7.4.1. Model architecture

To describe the tracking-based cross-correlation (CCG), our model adopts a linear cascade framework, in which the overall response is modeled as the convolution of multiple temporal filters arranged in sequence. Rather than fitting each processing stage independently, we focus on modeling the difference between the fixated background and dynamic background conditions. These two conditions share the same core processing components—low-level encoding (h1), position estimation (h2), and response

mapping (h_3)—but differ in whether an additional correlation-computation process is required. We represent the fixated background condition as a composite linear system, denoted $G_{Fix} = h_1 * h_2 * h_3$. For the dynamic background condition, we assume that an additional grouping process, h^* , which integrates temporal correlation signals, is inserted into the processing stream, resulting in the full system response expressed as $G_{Dyn} = h_1 * h^* * h_2 * h_3$. Since the convolution operation is commutative for linear time-invariant systems, this formulation can be algebraically rearranged as $G_{Dyn} = h_1 * h_2 * h_3 * h^* = G_{Fix} * h^*$. This expression allows us to isolate and estimate the additional temporal integration process associated with correlation-based grouping by comparing the dynamic background response to that of the fixated background, while holding all other processing components constant.

Figure 7-5. Conceptual pipeline of doing tracking task with fixated or dynamic background.

Conceptual model of the tracking task under fixated and dynamic background conditions. In the fixated background condition (black arrows), the stimulus is encoded by a low-level temporal filter h_1 , followed by a position estimation stage h_2 , and finally mapped to a motor response via h_3 . In the dynamic background condition (red arrows), an additional correlation grouping process h^ is required to extract meaningful structure from the temporally varying input. This process feeds into the position estimation stage and adds delay to the overall response.*

To accomplish this decomposition task, we simultaneously fit the G_{Dyn} and G_{Fix} with following form, which all the kernel is the exponential family, staying consistent with previous studies (Bonnen et al., 2015). In the following, all the red bolded term is the free parameters in the model.

$$G_{Fix}(t) = \mathbf{A}_{Fix} \left(e^{-\frac{(\log(t) - \mu_{Fix,1})^2}{2\sigma_{Fix,1}^2}} - \omega e^{-\frac{(\log(t) - \mu_{Fix,2})^2}{2\sigma_{Fix,2}^2}} \right) \quad (7-10)$$

$$h^*(t) = \left(\mathbf{A}_{corr} e^{-\frac{t^2}{2\sigma_{corr}^2}} \right) \cdot \mathcal{S}(t) \quad (7-11)$$

$$\mathcal{S}(t) = \begin{cases} 1, & t \geq 0 \\ 0, & t < 0 \end{cases}$$

$$G_{Dyn} = G_{Fix} * h^* \quad (7-12)$$

In our model, the system response under the fixated background condition is described using a mixture of log-normal kernels. The log-normal form ensures causality, such that the response is strictly zero for all times prior to $t = 0$. The additional component of interest, $h^*(t)$, represents the temporal integration window specific to correlation computation. It is modeled as a Gaussian function centered at $t = 0$, and to enforce causality, it is multiplied by a step function that masks out any response before $t = 0$. The effective width of this integration window is proportional to the standard deviation parameter σ_{corr}^2 , with the half-height bandwidth given by $\sqrt{2 \ln(2)} \sigma_{corr}$. The response function for the dynamic background condition, G_{Dyn} is then modeled as the convolution of the fixated background response G_{Fix} and the correlation-sensitive integration window h^* .

To extend the integration window framework to reaction time, we drew on the idea proposed by Ejima & Ohtani (1989), which interprets reaction time as the duration required for the system to accumulate processed sensory input until a decision threshold is reached. This accumulation can be modeled as the convolution of the stimulus with the system's impulse response. To account for variability in response time, we further incorporated the drift-diffusion model (Ratcliff, 2002), in which the accumulation process is driven by both a deterministic drift and stochastic noise, forming a drifting Brownian motion. The time at which the accumulated signal first crosses the threshold corresponds to the first-passage time of this stochastic process, which follows an inverse Gaussian distribution under standard assumptions (Folks & Chhikara, 1978). Based on this framework, we modeled the reaction time distribution in Experiment 5-1 using the following form:

$$P(T|\phi) = \frac{1}{\mathcal{A}} e^{-\frac{(\alpha - E(T|\phi))^2}{2\sigma_D^2 T}} \quad (7-13)$$

$$\mathcal{A} = \int_{T=0}^1 P(T|\phi) dT$$

Here, T denotes the reaction time, and ϕ represents the magnitude of within-area temporal correlation. The parameter α is the decision threshold that triggers a response, and is fixed at 1. Because the observed reaction times are constrained to fall within 1 second, we use a truncated version of the inverse Gaussian distribution, renormalized over the interval from 0 to 1 second. In this formulation, $E(T|\phi)$ represents the expected accumulated activation as a function of correlation strength ϕ , and is defined as follows:

$$E(T|\phi) = \mathbf{S} \cdot \int_{t=0}^T G_{Dyn}(t) \cdot \mathcal{M}(\phi) dt \quad (7-14)$$

$$\mathcal{M}(\phi) = 1 - e^{-\frac{\phi^2}{2\sigma_D^2}}$$

Here, the S is a scaling factor that controls the overall internal signal strength. The function $\mathcal{M}(\phi)$ modulates the relative signal strength based on the within-area temporal correlation ϕ , normalized with respect to the maximum value at $\phi = 1$, consistent with the condition observed in the continuous tracking task.

In the final part of the model, we describe the proportion of selection (i.e., registering a response) in the RT task as a function of correlation level:

$$P(\text{Selection}|\phi) = \int_{x=\tau}^{\infty} \mathcal{N}(x, |S \cdot \mathcal{M}(\phi), \epsilon_{Detect}^2) dx \quad (7-15)$$

This equation reflects a classical signal detection framework. The function $\mathcal{N}(x|\mu, \sigma^2)$ denotes the probability density of a normal distribution with mean $\mu = S \cdot \mathcal{M}(\phi)$, representing the maximal internal signal strength elicited by a stimulus with correlation ϕ , and variance ϵ_{Detect}^2 , which characterizes internal noise in the decision process. The parameter τ defines the response criterion, representing individual differences in detection threshold or bias. Since the RT task involves simple detection without explicit discrimination, such bias may strongly influence response probability, and is therefore explicitly modeled in this component.

Finally, all components of the data—including the tracking-based CCGs, the distribution of reaction times, and the proportion of selection—were jointly fitted using a maximum likelihood estimation (MLE) framework. The total log-likelihood function is the sum of three terms corresponding to the respective data types:

- Tracking data:

$$\begin{aligned}
& \ln L(A_{Fix}, \mu_{Fix \cdot 1}, \sigma_{Fix \cdot 1}^2, \omega, \mu_{Fix \cdot 2}, \sigma_{Fix \cdot 2}^2, A_{Corr}, \sigma_{Corr}^2 | CCG_{Fix}, CCG_{Dyn}) \\
&= \sum \ln \left(\mathcal{N} \left(CCG_{Fix}(t) | G_{Fix}(t), \text{var}(CCG_{Fix}(t)) \right) \right) \\
&+ \sum \ln \left(\mathcal{N} \left(CCG_{Dyn}(t) | G_{Dyn}(t), \text{var}(CCG_{Dyn}(t)) \right) \right)
\end{aligned} \tag{7-16}$$

Here, we assume that the correlation coefficient at each time point follows a normal distribution, with its mean determined by the model prediction and its variance empirically estimated from the data.

- RT data:

$$\ln L(\sigma_D^2, S, \epsilon^2 | RT(\phi)) = \sum_{\phi} \ln(P(RT|\phi)) \tag{7-17}$$

- Proportion of selection:

$$\begin{aligned}
& \ln L(\epsilon_{Detect}^2 | Selection(\phi)) \\
&= \sum_{\phi} \ln \left(\binom{N}{Selection(\phi)} \right. \\
&\quad \cdot P(Selection|\phi)^{Selection} \\
&\quad \cdot (1 - P(Selection|\phi))^{N-Selection} \left. \right)
\end{aligned} \tag{7-18}$$

7.4.2. Result

We present the model fitting results for the tracking task in Figure 7-6 and for the reaction time (RT) task in Figure 7-7. The model qualitatively fine describing both the fixated and dynamic conditions' tracking task, and in correspondence, we extract the integration windows in respective to different attributes and participants. This linear cascade model can additionally extend to reaction time data by adopting a non-linear version drift diffusion framework. The model performed well both on predicting the distribution and central tendency of the reaction time.

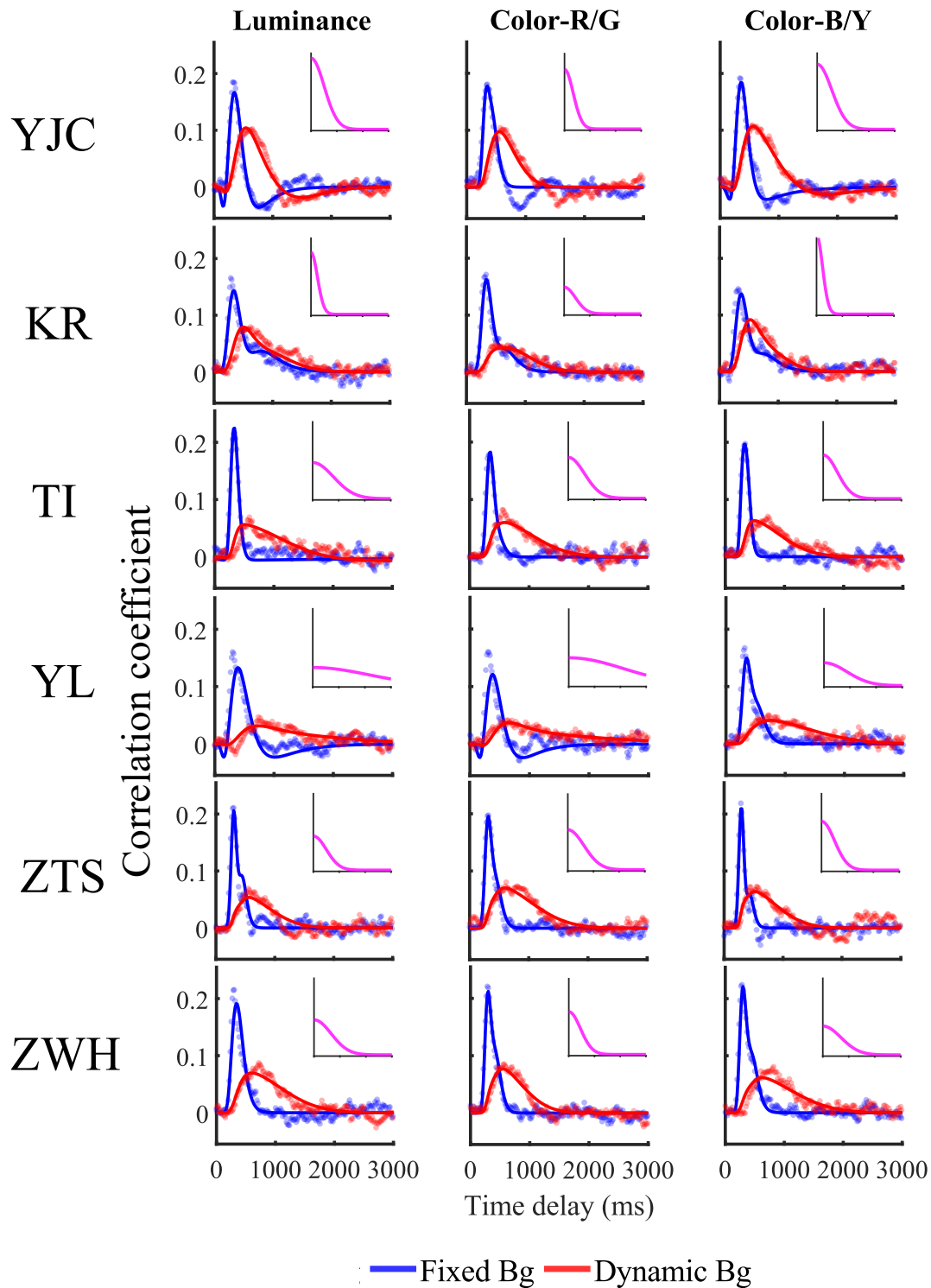


Figure 7-6. Model fitting result for tracking task for each participant and attribute types. Model fitting results for each participant and attribute type. Each panel shows the empirical cross-correlation data (dots) and corresponding model predictions (solid lines) for the fixed background (blue) and dynamic background (red) conditions. Data are organized by participant (rows) and stimulus attribute (columns: luminance, red-green, and blue-yellow). Insets in each panel display the estimated temporal integration window $h^*(t)$, reflecting the additional accumulation process required under dynamic background conditions.

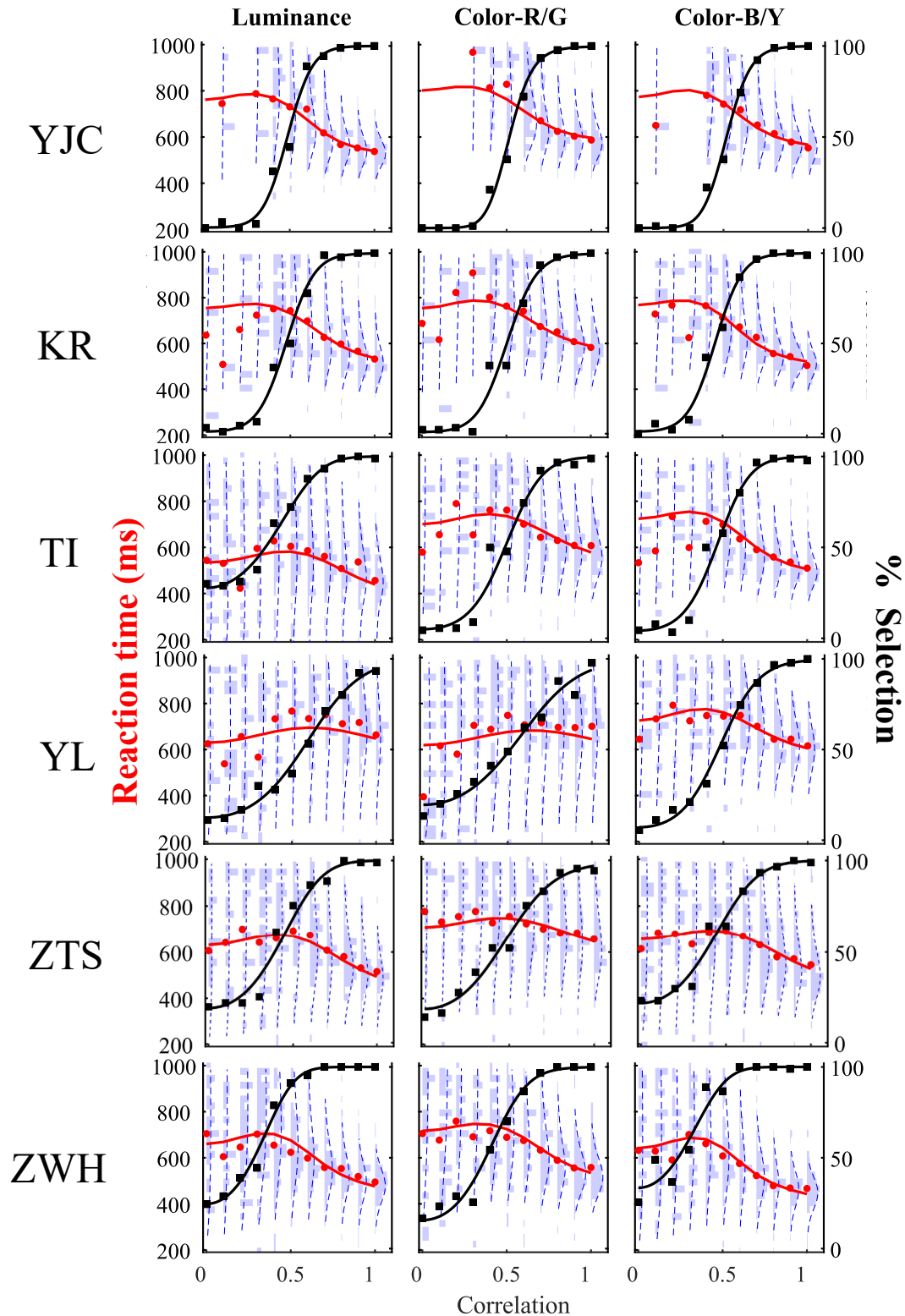


Figure 7-7. Model fitting result for RT task for each participant and attribute types.

Each panel shows individual results from the reaction time experiment, organized by participant (rows) and visual attribute (columns). Red dots indicate mean RTs from behavioral data; black squares indicate the observed proportion of selection. Blue histograms show the empirical RT distribution across trials, while blue dashed curves represent the predicted RT distribution based on the model. The red solid line indicates the predicted mean of the RT distribution, and the black solid line shows the model-predicted selection rate as a function of temporal correlation.

To further quantify the temporal characteristics of the estimated integration window, we computed the half-height bandwidth, defined as the time interval where the function value attenuates to the half value of the maximum. Summary statistics are presented in Figure 7-8. The results reveal individual variability in temporal integration, with one participant (YL) showing unreliable performance and model estimates that converged near the boundary of the parameter space. Therefore, this participant was excluded from group-level averaging. Across the remaining participants, the estimated integration windows were consistently within the range of approximately 600 to 1000 milliseconds, and no systematic differences were observed across the three visual attribute types.

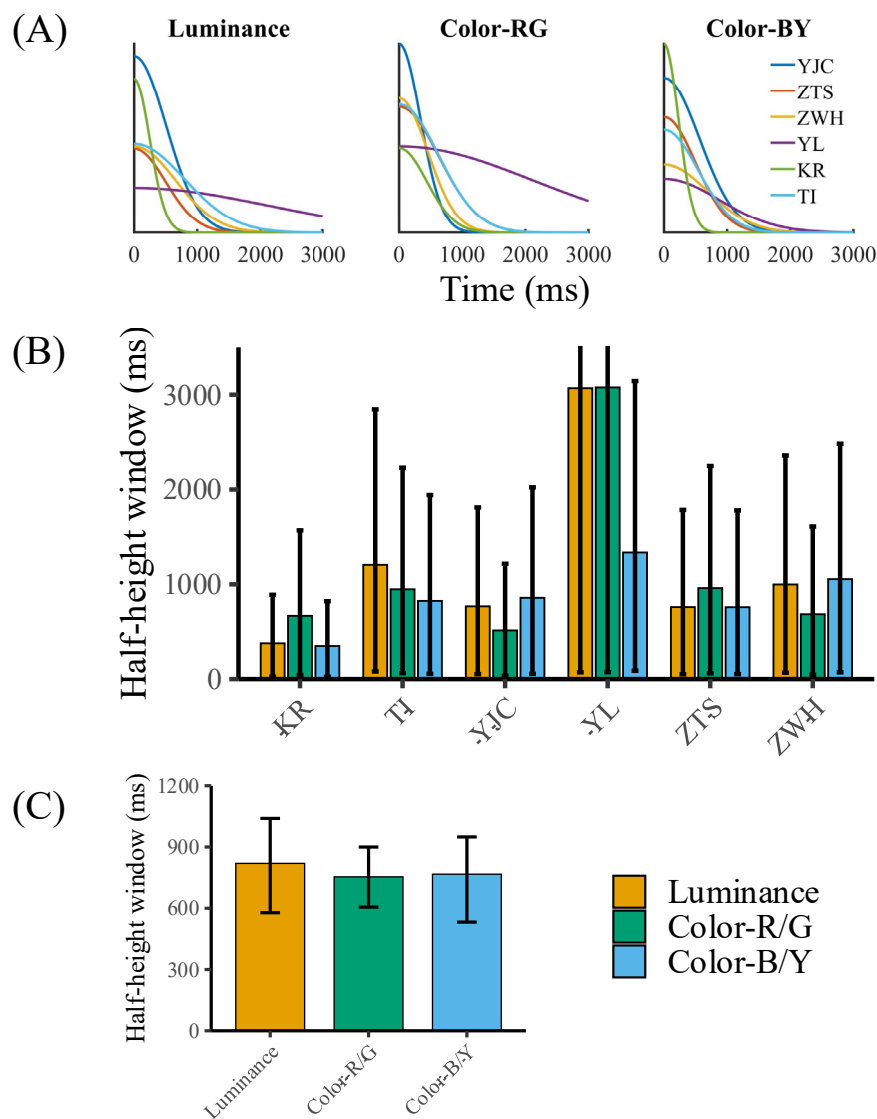


Figure 7-8. integration window overview and half-height bandwidth statistics.

7.5. General discussion

In this study, we aimed to determine the temporal extent over which the visual system integrates information to compute temporal correlation for spatial segmentation. As a preliminary step, we first examined whether the system truly utilizes temporal correlation as a relevant signal to perform the task. By measuring reaction times in response to stimuli with varying levels of temporal correlation, we observed that when the correlation level exceeded approximately 0.5, the system responded more rapidly as the correlation increased. This finding indicates that the visual system considers temporal correlation as a meaningful cue, rather than relying solely on frame-to-frame temporal changes. Based on this observation, we further investigated how temporal correlation influences the dynamic properties of tracking behavior. The results showed that the latency for detecting targets defined by spatial contrast and temporal contrast was comparable. However, when the target was defined by temporal correlation, the detection latency increased substantially, by nearly a factor of two. To characterize this additional computational demand, we employed a linear cascade model to estimate the underlying temporal integration process. The estimated integration window was consistently in the range of 600 to 1000 milliseconds across different participants and visual attributes. This suggests that the computation of temporal correlation involves a relatively slow but attribute-invariant process within the visual system.

7.5.1. Feature invariant properties of the temporal correlation-modulated process

Several attribute-dependent differences were observed in our study, and these patterns were consistent across both the reaction time and tracking tasks. However, this attribute-dependent is well captured by the low-level detector temporal characteristics, such as, stimuli defined by luminance changes elicited faster responses compared to those defined by red–green color modulation. This result aligns with previous findings showing that temporal variation is detected more rapidly in achromatic stimuli than in chromatic stimuli (Burr & Morrone, 1993; O’Donell & Colombo, 2008; Uchikawa & Yoshizawa, 1993). Based on this trend, one might expect that stimuli defined by blue–yellow modulation would result in even slower responses, since the S-cone pathway is known to have slower temporal processing compared to other visual channels (Shinomori & Werner, 2008; Umino et al., 2008). However, in our study, responses to blue–

yellow stimuli were as fast as those to luminance-defined stimuli. This result may reflect a limitation in the stimulus design, which did not strictly isolate the influence of S-cone mechanisms. It is possible that residual luminance artifacts were present in the blue–yellow stimuli, effectively masking the slower dynamics typically associated with S-cone processing. Despite this limitation, the consistent delay observed for red–green stimuli compared to luminance-defined stimuli suggests the involvement of attribute-dependent temporal mechanisms

Another attribute-dependent difference was observed in the peak latency of the fixated and plain background tracking conditions, where the red–green stimuli elicited consistently slower responses than luminance-defined stimuli. Two independent lines of evidence support this observation. First, Chen et al. (2024) demonstrated that the peak latency in continuous tracking closely matches the detection latency for local temporal variations. This suggests that the delayed responses in the red–green condition may be attributed to slower temporal processing in chromatic channels, consistent with our findings in the reaction time experiment. Second, the primary cue under the plain background condition arises from spatial contrast differences between consecutive frames, whereas under the fixated background, the cue is defined by temporal contrast at the same spatial location. Both types of cues are categorized as second-order motion signals. Prior studies have shown that chromatic second-order motion is processed more slowly than its achromatic counterpart (Garcia-Suarez & Mullen, 2013). Furthermore, color-defined motion is believed to engage additional or higher-order motion mechanisms compared to luminance-defined motion, possibly involving distinct motion parsers or even a third-order motion system (Cropper, 2005; Cropper & Derrington, 1994). The lack of a substantial latency difference between the fixated and plain background conditions in our data may be explained by the fact that both spatial contrast and temporal contrast represent second-order motion signals. Previous work has shown that these two types of second-order signals do not differ significantly in terms of processing latency (Ledgeway & Hutchinson, 2008; Miyamoto et al., 2020).

Despite the attribute-dependent differences observed in our results, the current study highlights a more fundamental attribute-invariant mechanism involved in temporal correlation processing. Evidence from Experiment 5-2 and computational modeling suggests the existence of a common integration window that remains consistent across different visual attributes. This integration process appears to introduce

additional processing time, likely reflecting the computation required to assess the temporal relationship between pairs of elements, such as determining whether their dynamics are temporally synchronous or asynchronous. This process contributes directly to segmentation performance. Previous studies have proposed the existence of a temporal simultaneity detector that is largely invariant to stimulus attributes, showing consistent performance across luminance, color, and motion (Fujisaki & Nishida, 2010; Nishida & Johnston, 2002) , and even across sensory modalities such as vision and audition (Fujisaki & Nishida, 2007; Parise & Ernst, 2016). This type of detector operates at a relatively low temporal frequency range, approximately 1 to 4 Hz, which is considerably slower than the low-level detectors that track local temporal variations (for a review, Holcombe, 2009). Although our study did not directly employ tasks such as temporal order judgment or synchrony detection, the segmentation paradigm used here also revealed attribute-invariant effects, in line with prior findings (Chen et al., 2024a; Guttman et al., 2005). Additionally, we observed notably long reaction times and broad integration windows, consistent with the low temporal resolution associated with attribute-invariant temporal segmentation (Chen et al., 2024a). Taken together, our findings suggest that visual segmentation based on temporal correlation may be driven, at least in part, by the output of an attribute-invariant simultaneity detector. This pathway, while consistent across stimulus types, operates with a relatively slow temporal profile, highlighting its distinct computational role within the broader visual system.

7.5.2. Limitation

In the reaction time results, we observed a slight upward trend in response latency under conditions with low temporal correlation, which may fall below the threshold required for reliably detecting temporally defined spatial structures. Although the selection rate in these low-correlation conditions lacks stability and interpretability, this trend may be partially attributed to participants' temporal expectations regarding stimulus onset. In our design, the onset of the structured stimulus followed a uniformly distributed foreperiod interval. Several studies have shown that the distribution and duration of the foreperiod can significantly influence measured reaction times (Han & Proctor, 2022; Houshmand Chatroudi et al., 2024; Karlin, 1959). It is therefore plausible that our measurements reflect not only the processing time for the stimulus but also an anticipatory component driven by the timing distribution. While the relative

differences between correlation levels remain valid due to the fixed interval distribution across all trials, the absolute values may be biased by expectancy effects. Future studies may consider whether similar anticipatory influences are also present in continuous tracking tasks, as they are in traditional reaction time paradigms.

Another limitation of this study concerns the assumption that the integration window for temporal correlation computation remains constant across different levels of correlation. This assumption simplifies the computational framework and attributes the increase in reaction time observed under lower correlation conditions to a prolonged decision process, required to accumulate weaker correlation signals. While this assumption is reasonable in the context of the tracking task—where only a single amplitude of correlation is modulated—it may not hold as readily in the reaction time task. Numerous studies on low-level sensory encoding have suggested that the visual system may adapt its temporal integration window depending on signal strength. For instance, luminance-based effects such as the Hess effect (Williams & Lit, 1983) and prolonged integration for low contrast stimuli (Fredericksen & Hess, 1997), indicate that weaker signals often lead to an extended integration period. It is therefore plausible that the temporal window for correlation computation also varies with signal strength. Future research should address this possibility and further investigate how integration dynamics adjust across the spectrum of temporal correlation, particularly in tasks requiring discrete responses.

A related limitation concerns the ambiguity of the underlying mechanism that contribute to the observed results. In our current model, the correlation-sensitive mechanism is represented by a single temporal kernel. However, in practice, this kernel may reflect a combination of multiple processes that all contribute to the computation of temporal correlation-based segmentation. For instance, previous studies have highlighted the role of recurrent processing in refining visual segmentation, particularly near object boundaries (Neumann & Sepp, 1999; Roelfsema et al., 2002), as well as the gradual propagation of attention within object contours through a "growing cone" mechanism (Jeurissen et al., 2016). It is likely that our measurements reflect a composite outcome of such processes, rather than isolating any one of them. Similarly, the current study does not address the specific neural implementations of these processes. The temporal profile captured here may correspond to computational dynamics that are distributed across multiple stages or circuits in the brain. Future work should aim to

disentangle these components with greater temporal precision and examine how distinct computational strategies contribute to correlation-based segmentation at the neural level.

7.5.3. Conclusion

This study investigated the temporal integration process underlying visual segmentation based on temporal correlation. Behavioral and modeling results revealed that segmentation performance improves with increasing correlation, and that detecting correlation-defined structure requires longer processing time than contrast-defined cues. The estimated integration window was consistent across attributes, spanning approximately 600 to 1000 milliseconds. These findings support the existence of a slow, attribute-invariant mechanism that underlies temporal correlation computation, extending the concept of simultaneity detection to the domain of visual organization.

Chapter 8. General discussion

To start the discussion of this thesis, we first recap the core research questions addressed throughout the previous chapters. The overarching aim of this work is to determine whether a higher-order, attribute-invariant mechanism exists that supports temporal segmentation by computing extended temporal correspondences and comparing dynamic signals across visual attributes. In addition to identifying its functional role, we also aim to characterize its computational principle, which we hypothesize to involve the integration of multiple temporal channels followed by correlation-based evaluation of temporal similarity. Through this operation, local dynamic signals are transformed into a structured representation that captures relational information across space and time, enabling segmentation that is invariant to the specific visual attribute from which the signal originated.

Synthesizing the findings across chapters, we propose a dual-stage framework for temporal segmentation, illustrated in Figure 8-1. This model incorporates both rapid, feature-dependent processing and a slower, feature-invariant integration stage that abstracts temporal similarity into a graph-like signal structure. The following sections provide a detailed discussion of how this framework accounts for the observed behavioral patterns, its computational implications, and directions for future research.

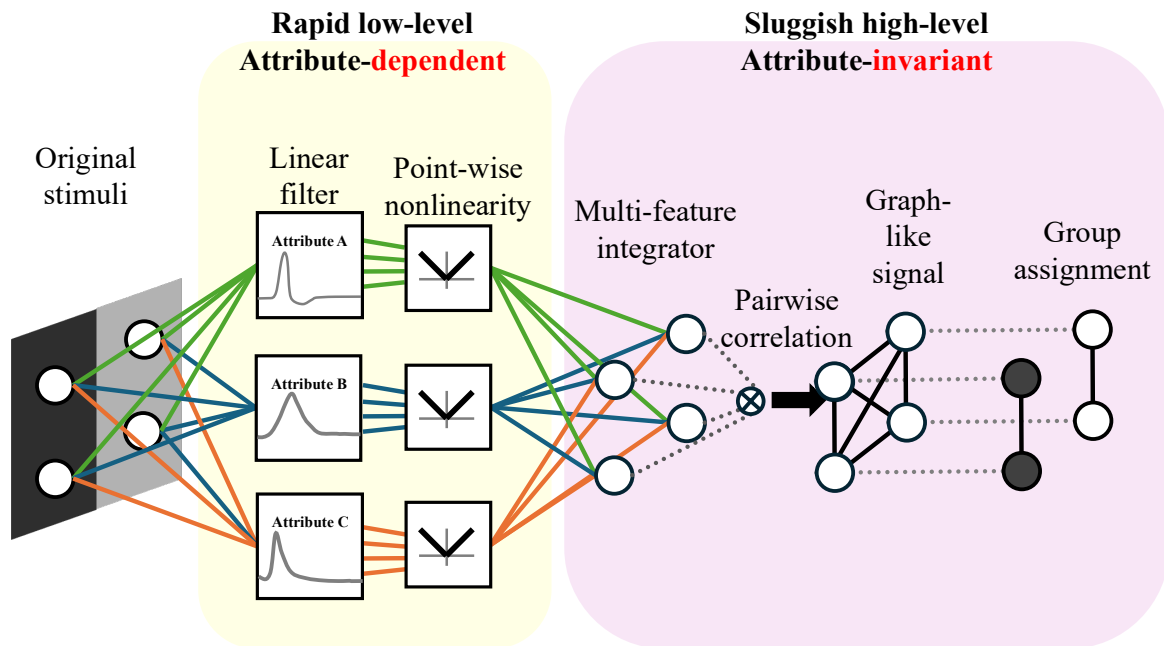


Figure 8-1. Conceptual framework integrating all result in the thesis.

This schematic illustrates the computational framework developed across Chapters 3–7. The early stage of processing (yellow region) consists of multiple low-level, attribute-specific pathways, each composed of a linear filter followed by point-wise nonlinearity, simulating the encoding of temporal signals for individual visual features (e.g., luminance, color, motion direction). These attribute-dependent signals are transmitted to a higher-order, sluggish processing stage (purple region), which integrates responses across features. The integration stage performs pairwise correlation computations among all spatial elements, yielding a graph-like signal that represents temporal similarity over space and across attributes. This structured representation supports the final group assignment process, which segments the input into coherent regions. The proposed architecture explains how the visual system can perform segmentation based on temporal correlation in a feature-invariant manner, and unifies the empirical and modeling results presented throughout the thesis.

8.1. Feature-invariant process of temporal segmentation

We first examined whether a feature-invariant process is involved in temporal-based segmentation by extending the tested visual attributes to include luminance, color, and motion direction. These three attributes were chosen because they exhibit markedly different temporal dynamics during low-level encoding, as illustrated in the yellow stage of Figure 8-1. If a higher-order, attribute-invariant mechanism exists, we would expect that, despite the heterogeneity in early encoding, the subsequent integration stage would impose a unified temporal resolution on the task. In other words, the system should exhibit consistent segmentation performance across different attributes if the signals are routed through the putative sluggish, integrative process. The results in Chapter 3 revealed both attribute-dependent and attribute-independent characteristics.

On the attribute-dependent side, the temporal resolution of segmentation varied systematically with the feature used: luminance supported the highest resolution, followed by color, and finally motion direction. On the attribute-independent side, when two different attributes were jointly involved in a segmentation task, the difference in temporal resolution across attributes was substantially reduced compared to single-attribute conditions. Moreover, the overall temporal resolution in these mixed conditions was comparable to or lower than that of the better-performing single-attribute case.

Beyond attribute-level effects, the most critical findings was that segmentation performance was strongly influenced by the temporal waveform of the signal itself. This suggests that the underlying comparison mechanism is sensitive not only to event timing but also to the spectral composition and shape of the raw temporal signal. Based on these observations, we propose that temporal correlation—quantified via Pearson correlation or cosine similarity—serves as a candidate computational principle. Such operations inherently preserve both frequency and phase information and thus offer a plausible mechanism for computing similarity across temporally varying signals.

To directly test this computational assumption that temporal correlation serves as the primary cue for temporal segmentation, we conducted a targeted test in Chapter 5. In this study, we directly manipulated the pairwise temporal cosine similarity among all elements within the stimuli. By carefully controlling both spatial and temporal low-level image statistics, we ensured that spatial structure could not be readily extracted from the stimuli, making temporal correlation the only available cue for segmentation. This design allowed us to isolate the influence of temporal correlation on visual segmentation, independent of specific visual attributes. The results supported our hypothesis: participants' segmentation performance systematically covaried with the temporal correlation structure of the stimuli. Notably, the cross-boundary correlation alone (i.e., correlation between elements across the segmentation boundary) was insufficient to account for the observed patterns. Instead, performance was highest when within-region temporal correlation was strong and cross-region correlation was weak. These findings suggest that the visual system evaluates a broad pattern of pairwise relationships rather than relying on a single contrast across a boundary. This implies that segmentation may be guided by a network- or graph-like representation of temporal similarity across the visual field, rather than by localized feature contrasts alone.

To further examine whether such graph-based computation suffices for segmentation, we employed a simulation approach. Specifically, we used a semi-image-computable model grounded in graph decomposition to assess whether such a structure could replicate human performance patterns. The simulation supported this possibility when a small set of free parameters was included to account for factors such as internal noise and distance-based constraints inherent to the human visual system. This model serves as an alternative to the traditional Filter-Rectify-Filter (FRF) framework, offering a complementary computational perspective on visual segmentation. Unlike FRF models, which rely on predefined spatial filters and are limited in handling interactions among temporally varying signals, the graph-based model enables structured computation over complex temporal dynamics. It formalizes segmentation as an operation on a similarity graph, allowing direct manipulation and testing of mid-level grouping processes based on temporal correlation. As such, it introduces a new, image-computable framework for investigating perceptual segmentation—one that expands the scope of testable conditions beyond what conventional models can accommodate.

Converging evidence also comes from behavioral response times from Experiment 1 in Chapter 7. In this experiment, we measured participants' reaction times in response to stimuli with varying levels of temporal correlation. The results showed that weaker temporal correlations were associated with longer detection times, consistent with the notion that weaker signals require more time to reach a decision threshold. This pattern supports the interpretation that the visual system actively computes temporal correlation—or a functionally equivalent metric—as part of the segmentation process.

Taken together, the findings presented in this section provide converging behavioral and computational evidence for a feature-invariant mechanism of temporal segmentation. By linking human performance to graph-based representations of temporal similarity, this work establishes a novel and generalizable computational framework that moves beyond traditional filtering-based models. This mechanism is illustrated in Figure 8-1. as the signal correlator and graph-based signal within the purple stage, which abstracts pairwise temporal relationships into a structured signal used for segmentation. As such, the proposed framework not only accounts for the observed perceptual patterns but also constitutes the core theoretical contribution of this thesis.

8.2. Integrated information processing stream

The underlying process of temporal segmentation may exhibit feature-invariant characteristics while still allowing for different modes of information integration. For example, the effective information stream could emerge from multiple parallel correlators, each operating on a distinct attribute channel and producing its own graph-like signal, which are then combined at a later stage. In this thesis, we further examined whether the visual system computes temporal correlation separately within multiple attribute-specific pathways—followed by a fusion step—or whether all attribute channels are first integrated into a common representational space before the correlation computation occurs. To test this, we used the quasi-white noise stimuli developed in Chapter 5, to examine cross-channel interference. Specifically, we investigated whether noise applied to one attribute channel could disrupt the temporal correlation computation in another, thereby revealing the extent to which the channels are processed independently or jointly.

The results showed that noise introduced in one attribute channel produced a comparable level of disturbance to the temporal correlation computation as noise originating from the same attribute. This suggests that information from different channels is merged, at least partially, prior to the computation of temporal correlation. If the channels were fully independent, noise in one should not affect the processing of another. To further characterize this interaction, we applied a simple linear model to describe how signals from different attributes combine. The model demonstrated that a linear combination of attribute channels could account for the majority of variance in the observed behavioral data. This provides additional support for the idea that diverse attribute signals are integrated into an amodal representation before temporal correlation is computed.

At this stage, we provide preliminary but foundational evidence suggesting that the feature-invariant characteristics of temporal segmentation can be attributed to an amodal integrator. This integrator appears to unify signals from different attribute channels into a shared representation prior to temporal correlation computation. The proposed multi-feature integrator is illustrated in Figure 8-1 as part of the overall architecture for feature-invariant segmentation.

8.3. Low temporal resolution and long receptive field on time

The final observation in our series of studies addresses a simple but fundamental question: if temporal correlation is indeed the key signal computed by the grouping and segmentation system, how much temporal information must be integrated to support this computation? To investigate this, we implemented a position tracking task across three stimulus conditions. In one condition, successful performance depended on computing the correlation between the temporal signals of the target and the background, while the other two conditions did not rely on such a computation.

The results showed that performance in the correlation-dependent condition was substantially slower than in the other two, suggesting that temporal correlation requires a longer processing duration. To further isolate the temporal dynamics of this computation, we applied a simple linear cascade model that separates the contribution of temporal correlation from the effects of early sensory encoding and motor output—components assumed to be shared across all conditions. The modeling results suggest that temporal integration spanning approximately 600 to 1000 milliseconds contributes meaningfully to segmentation performance. Importantly, this extended temporal window was consistent across different visual attributes, reinforcing the idea that the mechanism for computing temporal correlation operates independently of feature type.

Taken together, these findings support the notion that temporal segmentation relies on a sluggish, high-level mechanism that is both temporally extended and feature-invariant—consistent with the functional characteristics proposed at the outset of this work.

8.4. Some insight into the functional role of temporal segmentation

Perceptual grouping and segmentation are widely considered to play a critical role in organizing the inherently chaotic visual input into discrete perceptual units, thereby enabling both upstream abstraction and downstream interaction. One essential function is to delineate the effective area of interaction for an object, defining the spatial extent to which attentional, motor, or inferential processes should apply. Notably, perceptual grouping and segmentation can often be resolved from a single static snapshot. Contemporary computer vision models also reflect this view, demonstrating that high-fidelity segmentation can be achieved from individual frames alone (Segment Anything,

Jiang et al., 2022; YOLO review, Kirillov et al., 2023). This raises a fundamental question: why has the human visual system evolved a mechanism for temporal grouping and segmentation, if static information is often sufficient?

One possible explanation lies in the inherent instability and complexity of the optic field in natural environments. Most perceptual estimates of object properties—such as position, curvature, or depth—depend heavily on the pattern of light reflected from object surfaces. These estimates can be relatively reliable under idealized conditions, such as uniform illumination and diffusely reflecting surfaces. However, in many real-world situations—such as highways at night, disco halls with flickering lights, scenes with overlapping or multiple light sources, or environments involving transparent, glass-like, or specular surfaces—the optical signals become distorted, discontinuous, or temporally inconsistent. In such cases, segmentation based on a single snapshot becomes extremely fragile, often yielding perceptual instability across time. These observations suggest that integrating visual information over time may allow the system to form more stable and reliable representations of object structure. For instance, Sun et al. (2025) demonstrated that machine learning systems trained on datasets containing non-diffuse surfaces can spontaneously develop mechanisms for extracting long-timescale motion patterns to improve optical flow estimation. Such findings highlight the computational advantage of temporally extended processing, especially under conditions where spatial cues alone are insufficient.

Apart from signal instability arising from exogenous sources, the visual system must also contend with endogenous variability. Even under stable viewing conditions, the image projected onto the retina is inherently unstable due to continuous eye movements, including saccades, microsaccades, fixational drifts, and tremors. These movements constantly shift the retinal image, causing spatial overlap or displacement of signals from different objects. To maintain perceptual continuity, the visual system must develop mechanisms to resolve ambiguities introduced by eye movement-induced signal fusion and misalignment. Temporal correlation computations provide a compelling candidate for addressing this challenge. Even if signals from different objects transiently stimulate the same retinal neurons—or conversely, if signals from the same object activate different neurons over time—temporal coherence can serve as a statistical cue to rebind consistent features across time and segregate unrelated ones. By

tracking correlated temporal patterns across neural populations, the system can distinguish stable object representations from spurious overlap caused by self-motion.

From these perspectives, temporal segmentation may not simply be an auxiliary strategy, but a necessary computational adaptation that enables the visual system to maintain perceptual coherence in dynamic and ambiguous environments. By selectively integrating time-varying signals, the system can suppress transient noise and extract stable structural information. Thus, the ability to segment based on temporal correlation may reflect a core function aimed at enhancing perceptual stability under real-world constraints.

8.5. Limitation

Through these studies, we have identified several computational-level characteristics of how temporal correlation may serve as a basis for visual segmentation. In addition, we have proposed a conceptual workflow that outlines a plausible mechanism integrating these findings into a unified framework. Nevertheless, perceptual grouping and segmentation remain longstanding research domains, encompassing a wide range of phenomena and functional dimensions that extend beyond the scope of this thesis. Here, we organize the remaining limitations and future challenges into two major directions. The first concerns the development of image-computable models that are capable of operating in real-world conditions while remaining functionally aligned with human perceptual systems. The second involves the identification of neural substrates that can implement hypothesized computational architecture, including the temporal integration and cross-attribute grouping processes. In the following sections, I briefly outline these open issues and their potential implications.

8.5.1. A more realistic computational model in the natural world

While the current findings provide insight into the computational principles of temporal segmentation under controlled laboratory conditions, it remains unclear how such a mechanism operates in the complexity of natural scenes. Real-world visual environments are inherently noisy, dynamic, and often ambiguous, posing additional challenges to the reliability and stability of perceptual segmentation. To address this gap, a key future direction is to construct an image-computable model that instantiates the proposed temporal segmentation process and can operate directly on naturalistic

input. In the following subsections, we explore four core considerations toward this goal. Section 8.5.1.1 examines how temporal segmentation interacts with purely spatial grouping, and how these two processes may complement or compete with each other. Section 8.5.1.2 extends the discussion to more complex and heterogeneous visual features, considering whether the mechanism scales to naturalistic environments. Section 8.5.1.3 focuses on the logic of cross-attribute integration, elaborating on how signals from different visual domains might be temporally aligned and fused. Lastly, Section 8.5.1.4 discusses the need for flexible rules of shape extraction, especially under non-rigid or ill-defined object boundaries, which may require adapting segmentation strategies based on temporal regularities.

8.5.1.1. Generalizability to pure spatial segmentation and their interaction

In current theories of perceptual grouping and segmentation, spatial and temporal cues are often treated as distinct sources of information. The present work focuses primarily on temporal-cue-driven segmentation and thus cannot directly address whether a similar feature-invariant mechanism underlies spatial grouping processes. Nevertheless, a substantial body of evidence suggests that spatial and temporal cues interact in complex and inseparable ways (Fahle & Koch, 1995; S.-H. Lee & Blake, 2001; Motoyoshi, 2004; Parton et al., 2001; Rideaux et al., 2016). Disentangling these interactions is inherently difficult, as the perceptual system does not process space and time completely independently. For example, even when two spatially separated dots are presented simultaneously, differences in luminance magnitude can lead to temporal misalignment in perception—a phenomenon known as the Hess effect (Williams & Lit, 1983). This misalignment can, in turn, induce illusory depth perception, as in the Pulfrich effect (Lit & Hyman, 1951), despite the stimuli being physically located on the same depth plane. These findings illustrate how temporal information can be implicitly triggered by purely spatial configurations. Conversely, temporal modulation can also generate or refine spatial representations. For instance, studies have shown that eye movements, particularly fixational eye movements, can enhance spatial resolution and reveal latent spatial structure (Rucci & Victor, 2015; Wallis, 2005). These examples underscore the inseparability of spatial and temporal processing in natural vision. Future investigations should reconsider whether spatial and temporal segmentation rely

on entirely separate mechanisms, or whether they reflect different facets of a unified processing system. In addition, it remains an open question how spatiotemporal information is integrated or interacts within the broader architecture of visual grouping and segmentation.

8.5.1.2. The relationship with high-order spatial features

In the present study, we focused on visual segmentation based on relatively simple visual features such as luminance, color, and motion. Although these features span major domains of low-level visual processing, the visual system is also capable of extracting more complex structures, including contours, faces, and objects. For instance, research on ensemble perception has demonstrated that observers can estimate statistical properties of higher-order features—such as average size or average facial identity—across groups of items (for a review, Whitney & Leib, 2018). Such computations imply a spatial pooling process across distributed elements, which closely aligns with the computational goals of perceptual grouping and segmentation. Recent studies have begun to explore the relationship between ensemble perception and grouping mechanisms (Elias & Sweeny, 2020; Haberman & Whitney, 2011), but the precise interaction between these processes remains unclear. A related class of stimuli is the Glass pattern (Glass, 1969; Glass & Pérez, 1973), in which globally coherent structures emerge from locally paired elements. These patterns can be explained by models such as the Filter-Rectify-Filter (FRF) framework (Wilson & Wilkinson, 1998), yet it remains an open question whether such patterns themselves can act as building blocks for even higher-level spatial structure. More generally, one may ask whether spatial structure can be inferred from stimuli that are themselves already the outcome of lower-level grouping—such as perceiving a character “H” composed of local “A” elements. Both ensemble perception and Glass pattern stimuli raise a common theoretical question: although visual features are hierarchically organized, it is not yet clear whether—and how—the visual system infers new spatial structures from inputs that already reflect prior levels of grouping. Future work should investigate whether the mechanisms described in this thesis can be extended to account for such hierarchical organization in perceptual processing, and their spatiotemporal characteristics.

8.5.1.3. Multi-attribute integration rules

In this study, we investigated how different visual attribute channels contribute to temporal segmentation by simultaneously modulating two feature types within the same trial. This design allowed us to test the presence of attribute-invariant integration and revealed that a linear combination model could largely explain the observed performance. However, two important cases remain unaddressed, limiting our ability to fully characterize the rules governing multi-attribute integration. First, it remains unclear how the segmentation mechanism would operate when more than two attributes are jointly modulated in a single stimulus. While the observed invariance in our current experiments suggests the presence of an amodal signal extractor, the interactions between just two attributes may not be sufficient to reveal the full computational structure of this integrator. In particular, such two-way interactions may fail to capture potential nonlinearities or capacity limitations that only emerge in higher-dimensional feature spaces. Including three or more simultaneously varying attributes may exaggerate the interference effects, allowing for stronger tests of whether the integration process is truly generalizable or constrained by feature-specific saturation or prioritization mechanisms. Second, in all of our experimental designs, each spatial location was associated with a single attribute value—meaning that any given local region was defined by only one visual feature at a time. This setup, while simplifying the stimulus structure, does not reflect more naturalistic conditions in which multiple attributes co-exist and are spatially co-located. As such, the current design does not allow us to determine whether the segmentation mechanism can disentangle or integrate signals when multiple features occupy the same spatial region. This is particularly relevant for testing whether the amodal extractor treats all features equally when they are spatially overlapping, or whether competition and channel-specific gating occurs at the encoding stage. Future research may further explore the internal structure of the integrator and examine how different sources contribute to cross-attribute interactions.

8.5.1.4. The rule for arbitrary shape extraction/organization

To model the human segmentation performance, we adopted a graph-based segmentation model that offers theoretical flexibility to account for a wide variety of spatial structures. However, the shapes used in our experiments were relatively simple—primarily circles and squares—which are highly regular and symmetric. It remains

uncertain whether the model can achieve comparable performance when applied to more irregular or complex spatial forms. To address this limitation, future research should examine whether the proposed framework generalizes to a broader range of shape configurations. In addition, the spatial resolution of the segmentation mechanism—how precisely the system can localize and track boundaries defined by temporal correlation—remains untested. Even if a segmentation decision is made, it is unclear whether the inferred structure reflects coarse region grouping or fine-grained parsing. This may require new experimental methods to capture participants’ perceived segmentation boundaries with greater spatial precision. For example, recent work has proposed techniques for stimulus-wise estimation of subjective boundaries, which could provide a more detailed understanding of how observers parse complex visual input (e.g. Vacher et al., 2023). Such data would allow segmentation theories to move beyond identifying influencing factors and begin to characterize the computational processes by which perceptual structures are inferred.

8.5.2. Neurophysiological insight

We demonstrated that the visual system is highly sensitive to graph-like representations that encode complex temporal information within spatial configurations. However, the present work focuses exclusively on the computational-level description of these phenomena and does not provide direct insight into the underlying neural implementation. While we do not attempt to specify the exact neuronal computations involved, we aim to outline several directions through which future research may begin to bridge behavioral findings with neurophysiological mechanisms. Notably, the stimuli used in our experiments cannot be explained by traditional Filter-Rectify-Filter (FRF) models, which are often employed to simulate the response properties of complex cells. This suggests that the relevant processing may occur at stages beyond early visual cortex. Accordingly, future physiological studies may benefit from targeting higher-order areas to identify regions that exhibit selective responses to the kind of graph-like segmentation task examined in this work. Candidate regions may include areas along both the dorsal and ventral visual pathways—particularly the intraparietal sulcus (IPS) and inferotemporal cortex (IT)—which are known to support feature integration and abstract relational processing across visual domains. Another open question concerns how the nervous system supports sensitivity to graph-like structures at the behavioral level.

It remains unclear whether the brain explicitly performs computations that can be formally characterized as graph-based, or whether such sensitivity arises from the convergence of neural operations whose aggregate behavior resembles graph-theoretic principles. Distinguishing between these possibilities—explicit computation versus emergent approximation—will be crucial for developing biologically plausible models of mid-level segmentation.

8.6. Conclusion

In summary, this study employed a series of psychophysical experiments combined with computational modeling to investigate the visual system's ability to perform spatial segmentation based on temporal signals. We proposed the existence of a higher-order grouping and segmentation system that integrates temporally extended, cross-channel input and exhibits system-level behavior consistent with pointwise correlation computations structured as a graph. Based on the resulting graph representation, the system can directly infer spatial segmentation boundaries. Our findings suggest that this mechanism generalizes across several low- to mid-level visual features, including luminance, color, and spatial phase (as manifested in motion direction). While the current work focused on relatively simple stimuli and graph structures, future research should aim to explore the system's capacity for processing more complex graph signals, its ability to resolve spatiotemporal interactions, and the neurophysiological basis through which such computations are implemented. Advancing our understanding in these areas may offer new insights into the principles underlying mid-level vision in the human brain and contribute to a more comprehensive theory of perceptual organization.

Chapter 9. Appendix and supplementary material

9.1. Supplementary material of Chapter 3

Within vs cross attributes, Square waves

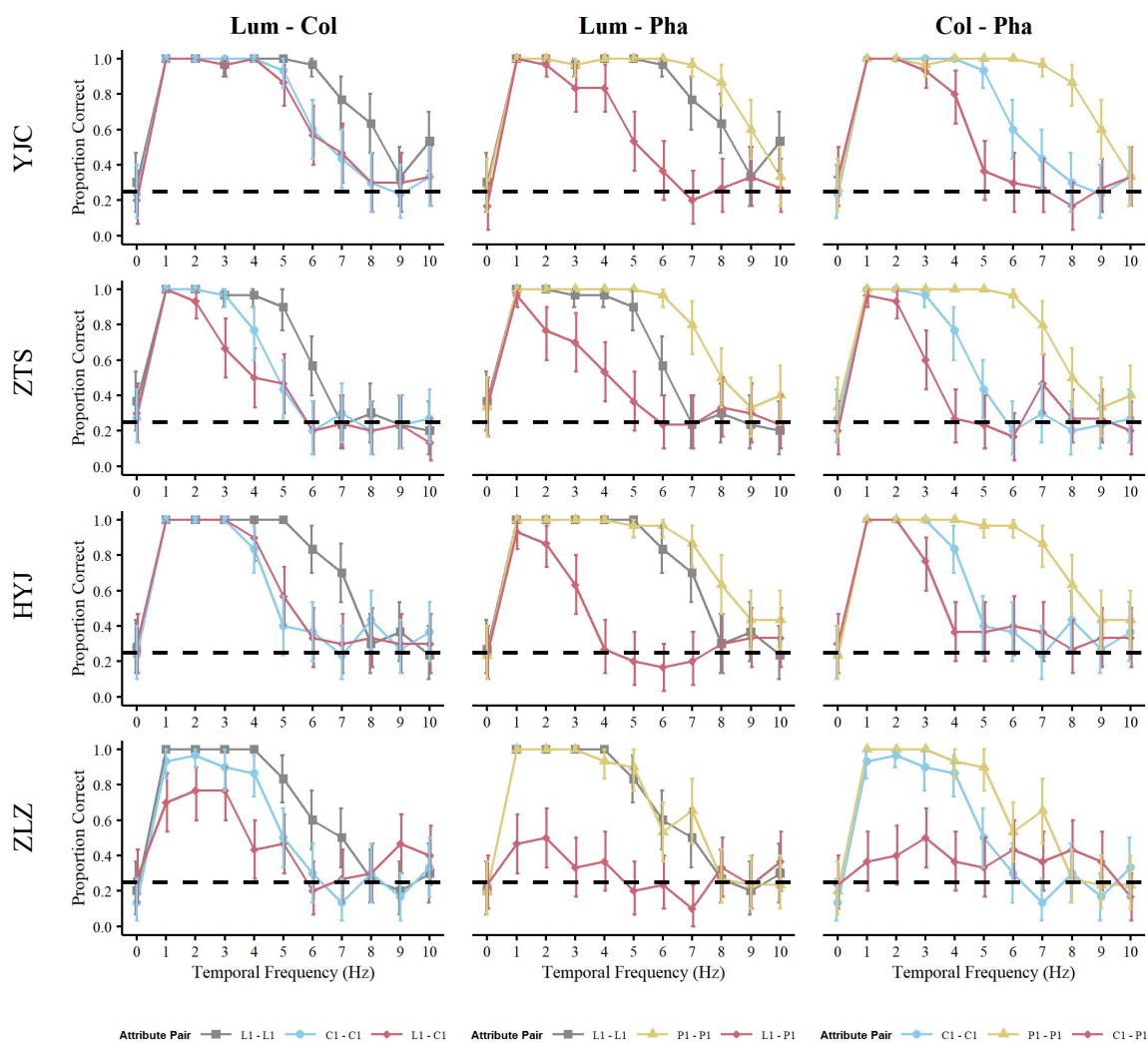


Figure 9-1. Individual result corresponds to the first row of Figure 3-5

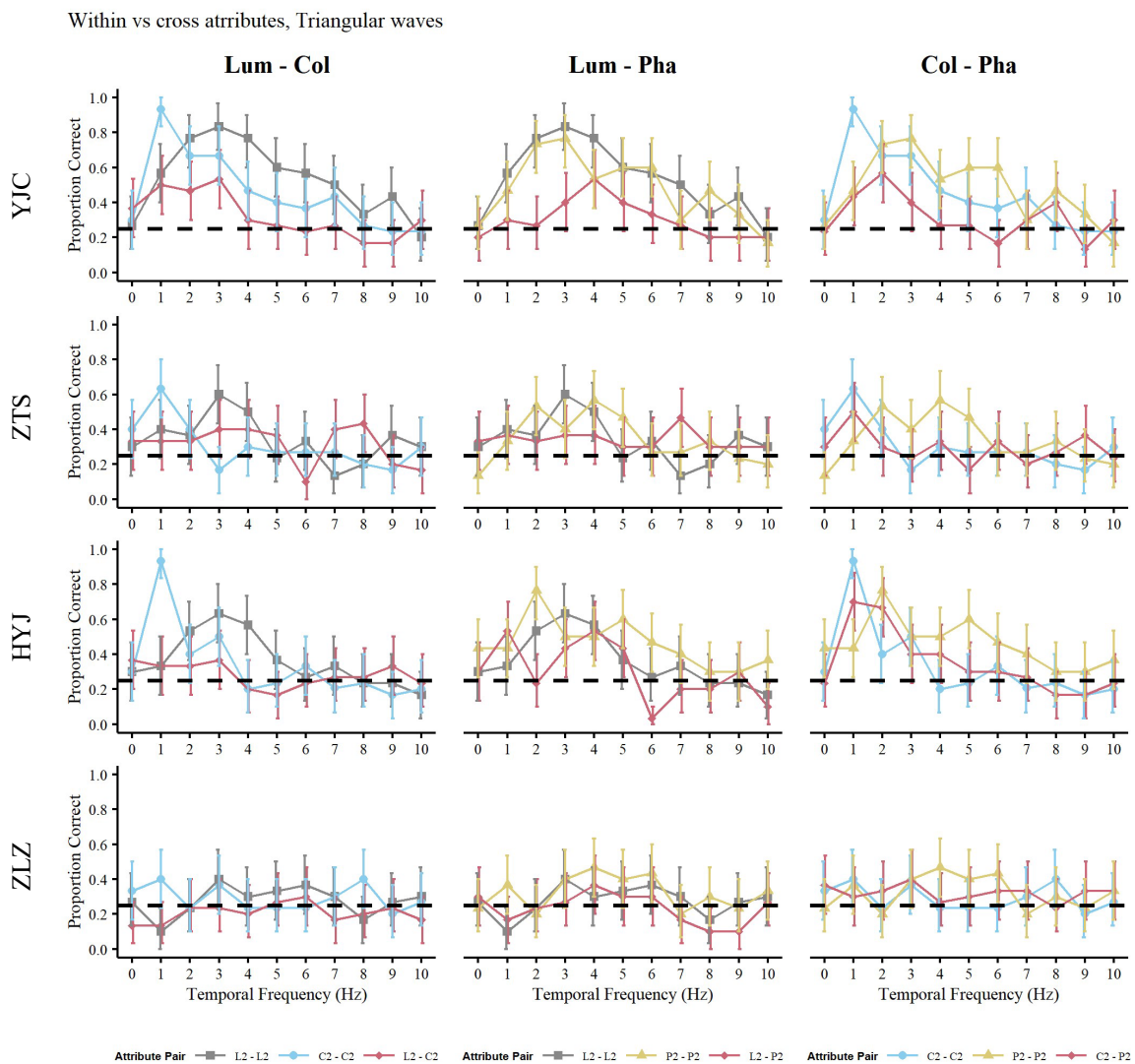


Figure 9-2. Individual result corresponds to the second row of Figure 3-5

Within-attribute, all waveform combinations

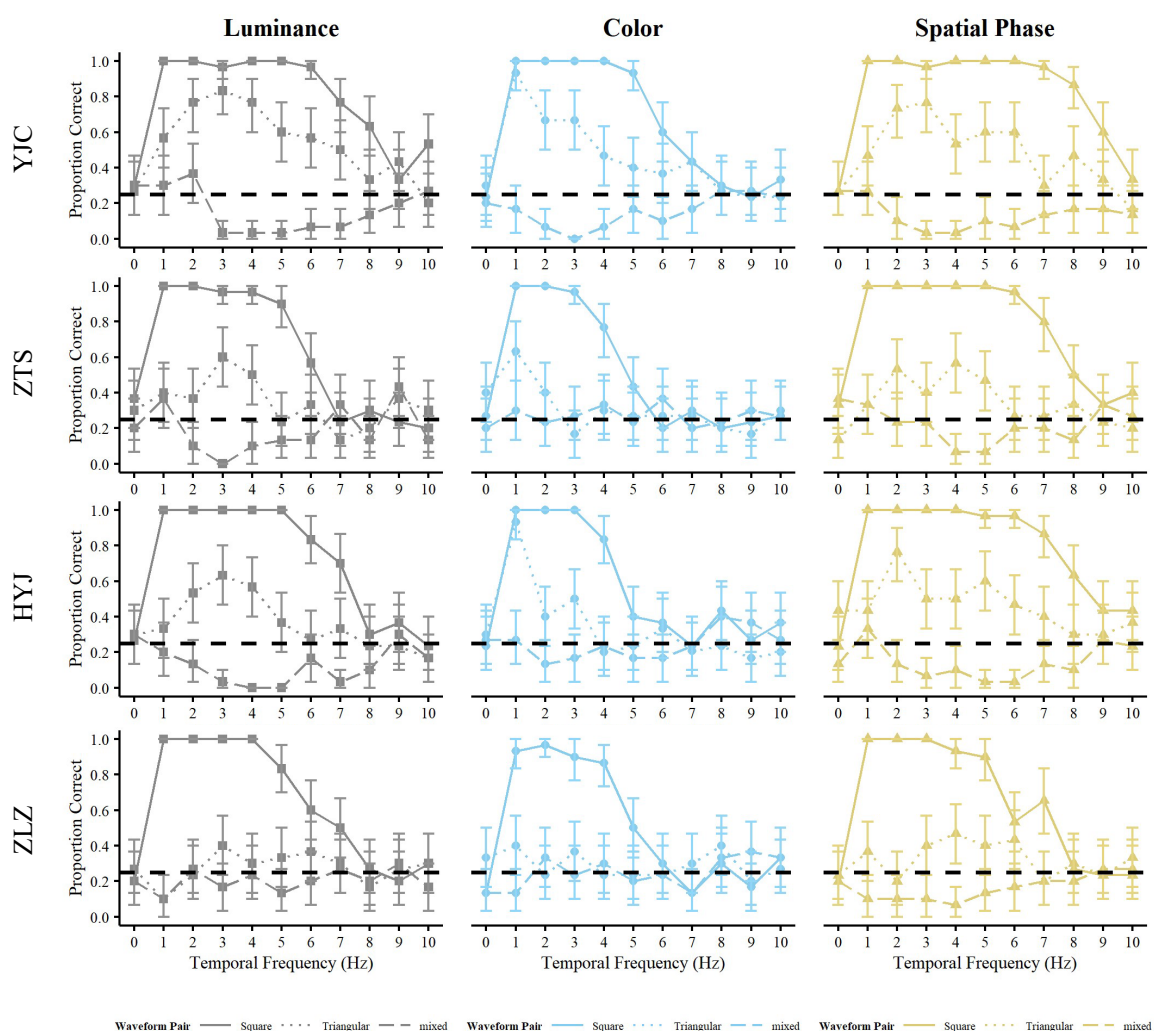


Figure 9-3. Individual result corresponds to the third row of Figure 3-5

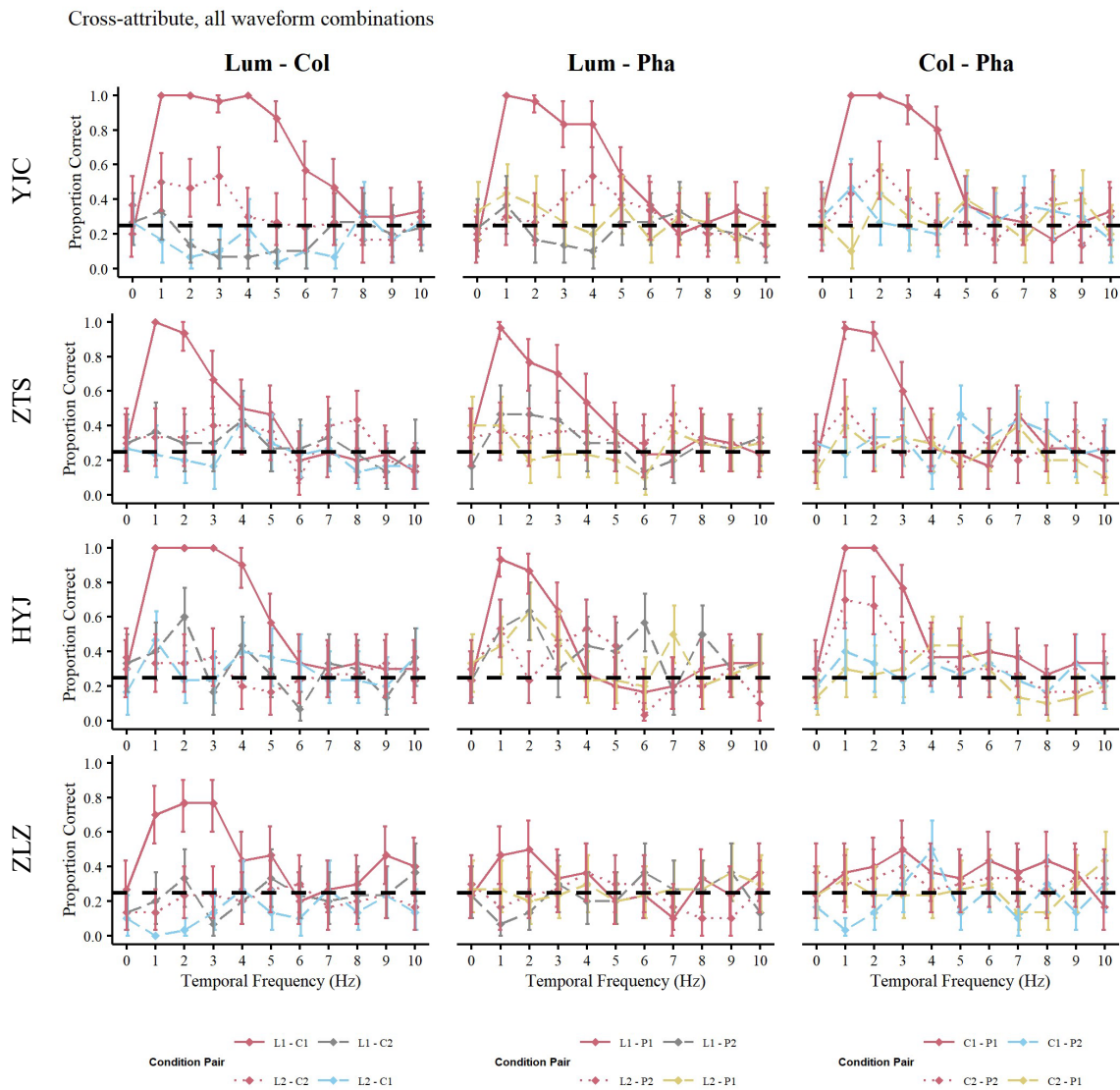


Figure 9-4. Individual result corresponds to the forth row of Figure 3-5

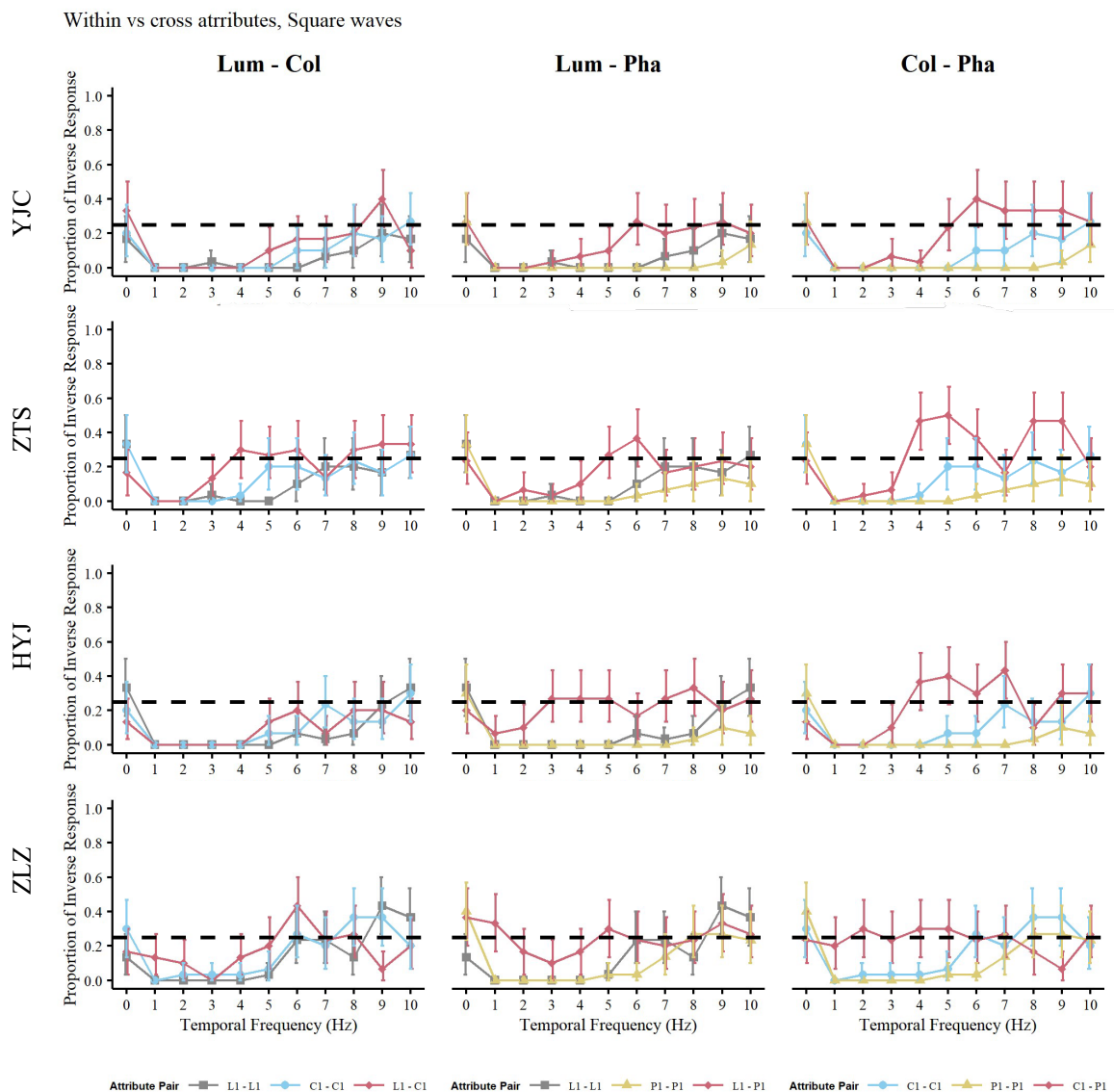


Figure 9-5. Individual result corresponds to the first row of Figure 3-6

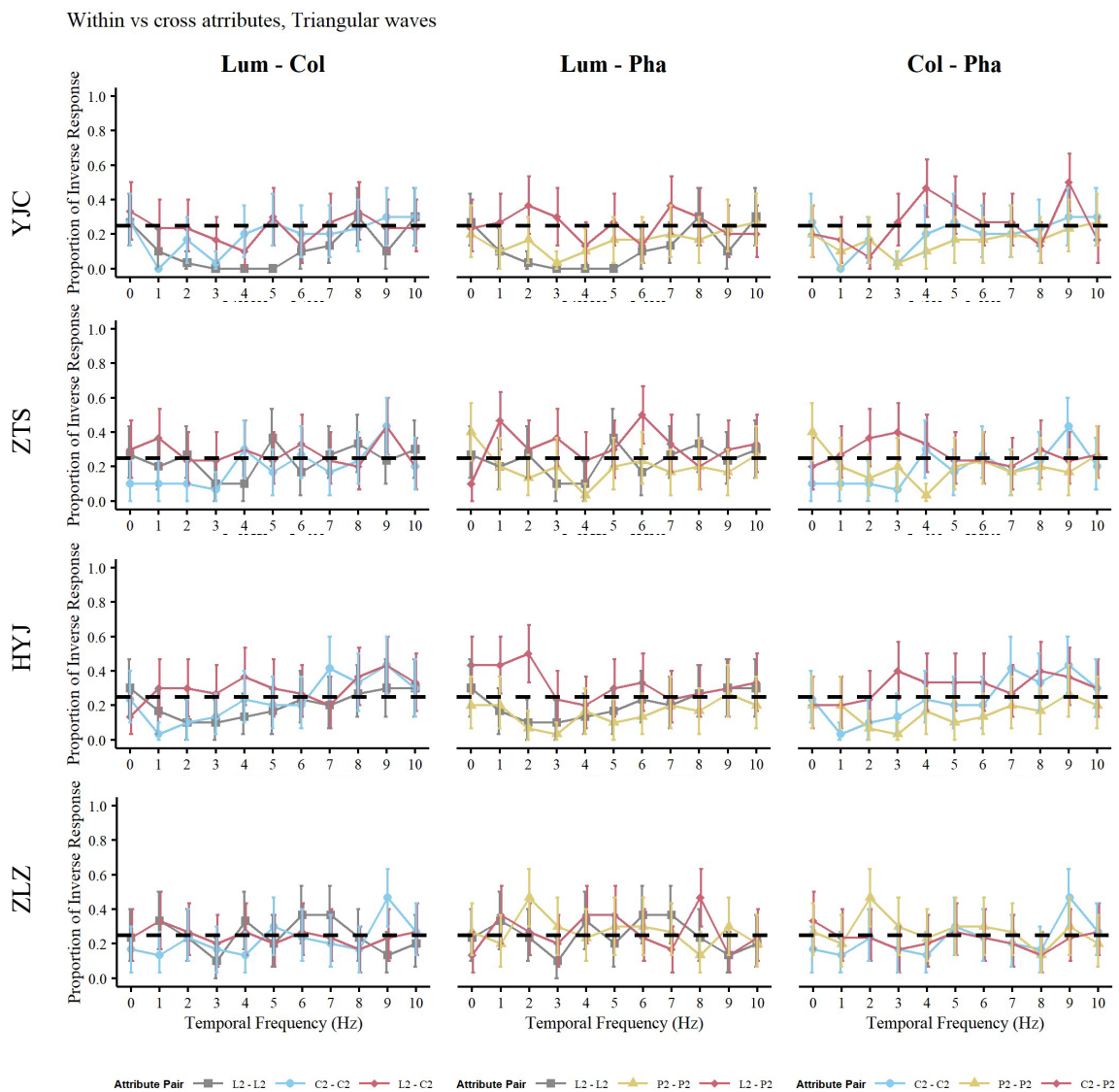


Figure 9-6. Individual result corresponds to the second row of Figure 3-6

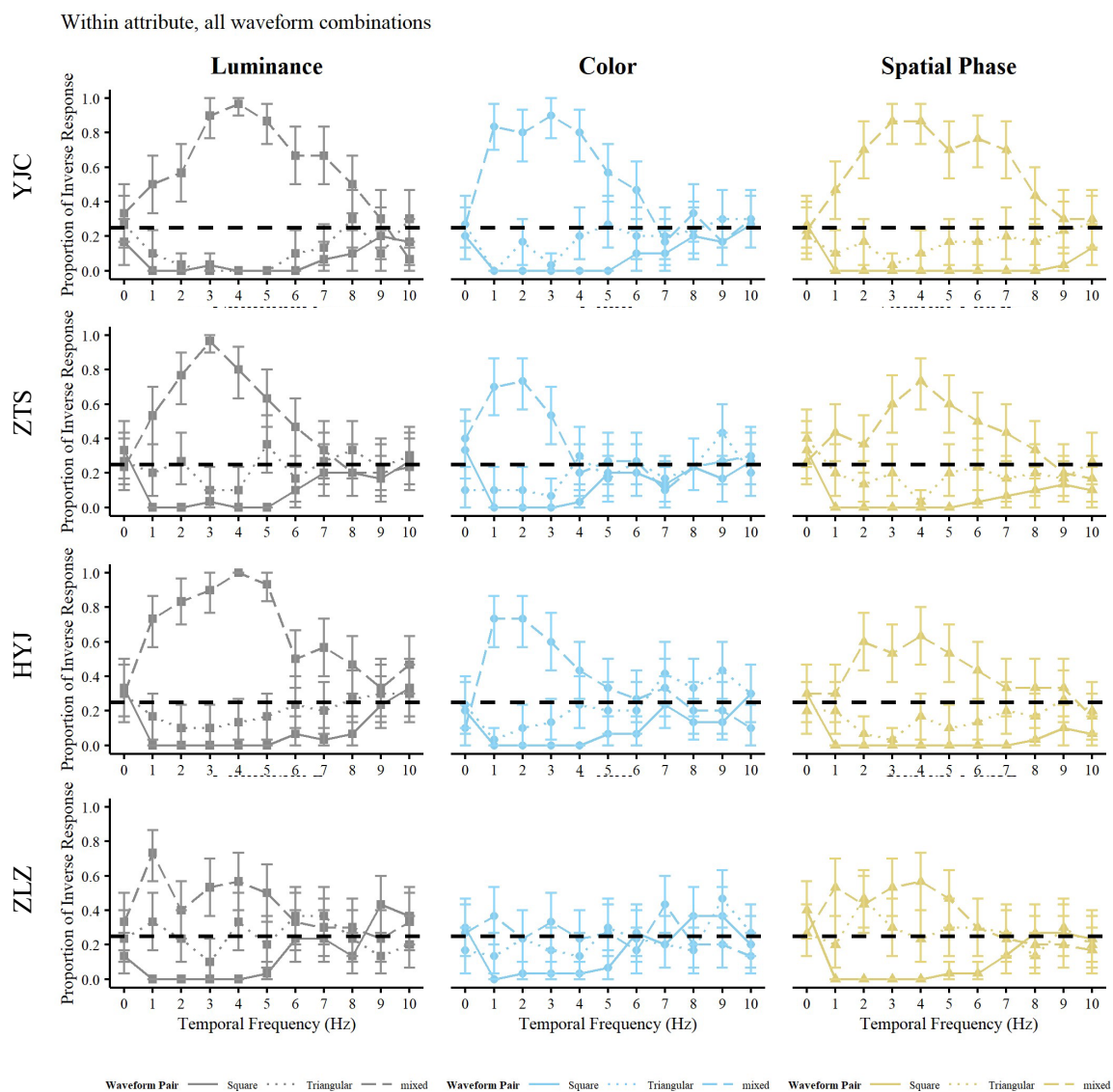


Figure 9-7. Individual result corresponds to the third row of Figure 3-6

Cross attribute, all waveform combinations.

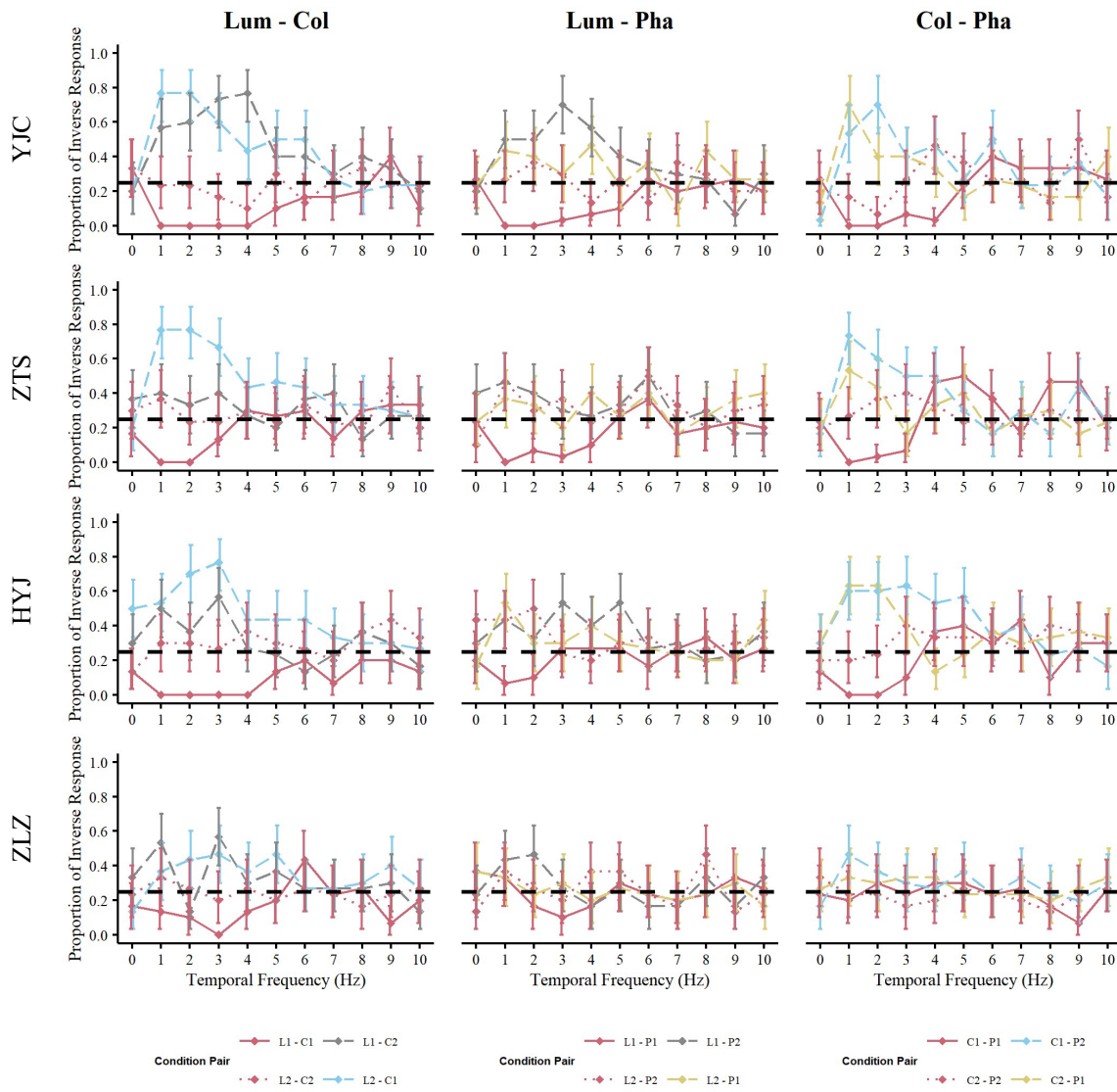


Figure 9-8. Individual result corresponds to the forth row of Figure 3-6

9.2. Supplementary material of Chapter 5

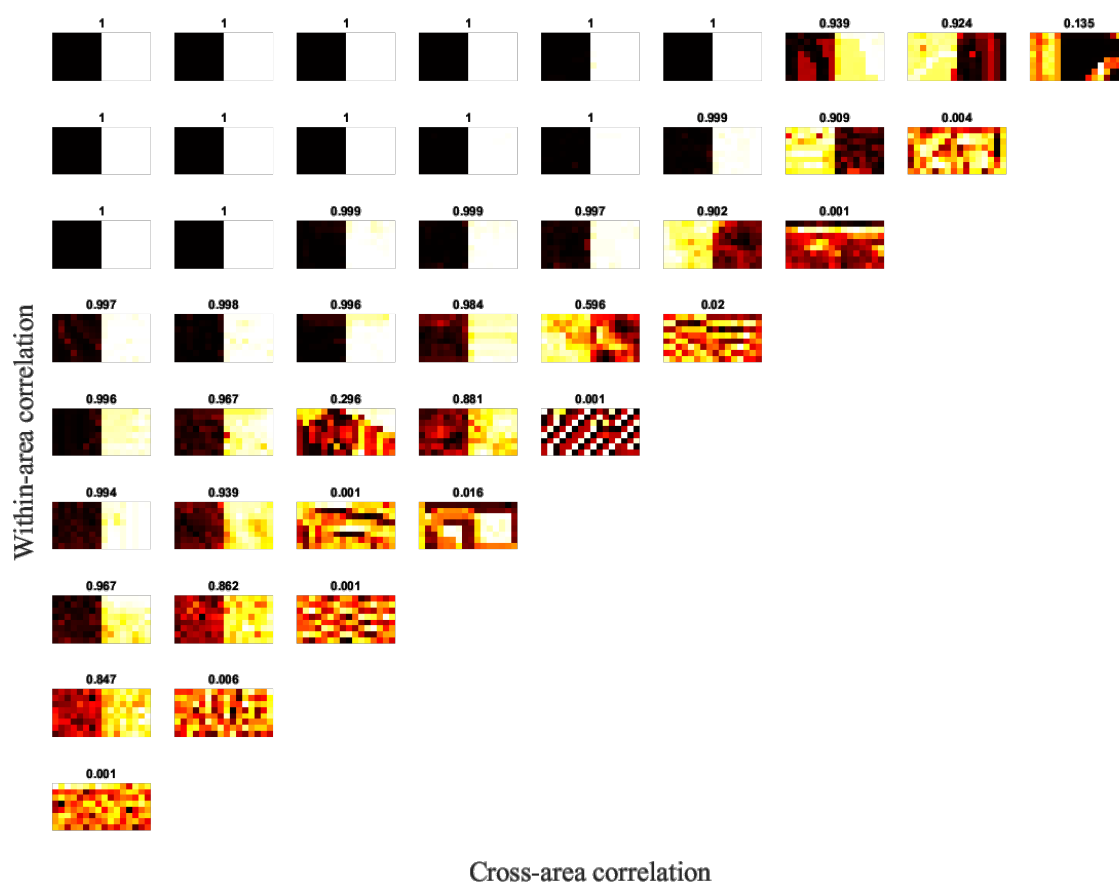


Figure 9-9. Single trial output of the naïve model for signal interval input

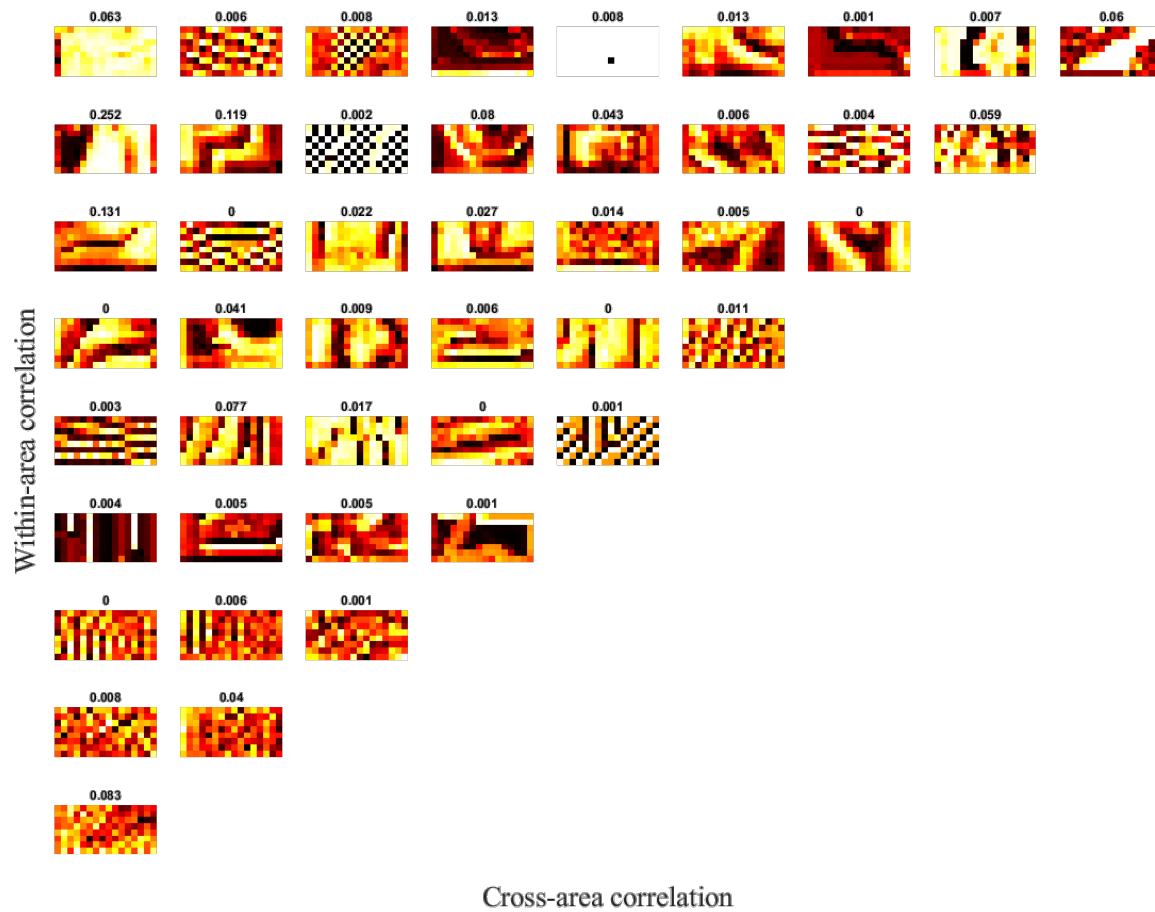


Figure 9-10. Single trial output of the naïve model for signal interval input

9.3.Supplementary material of Chapter 6

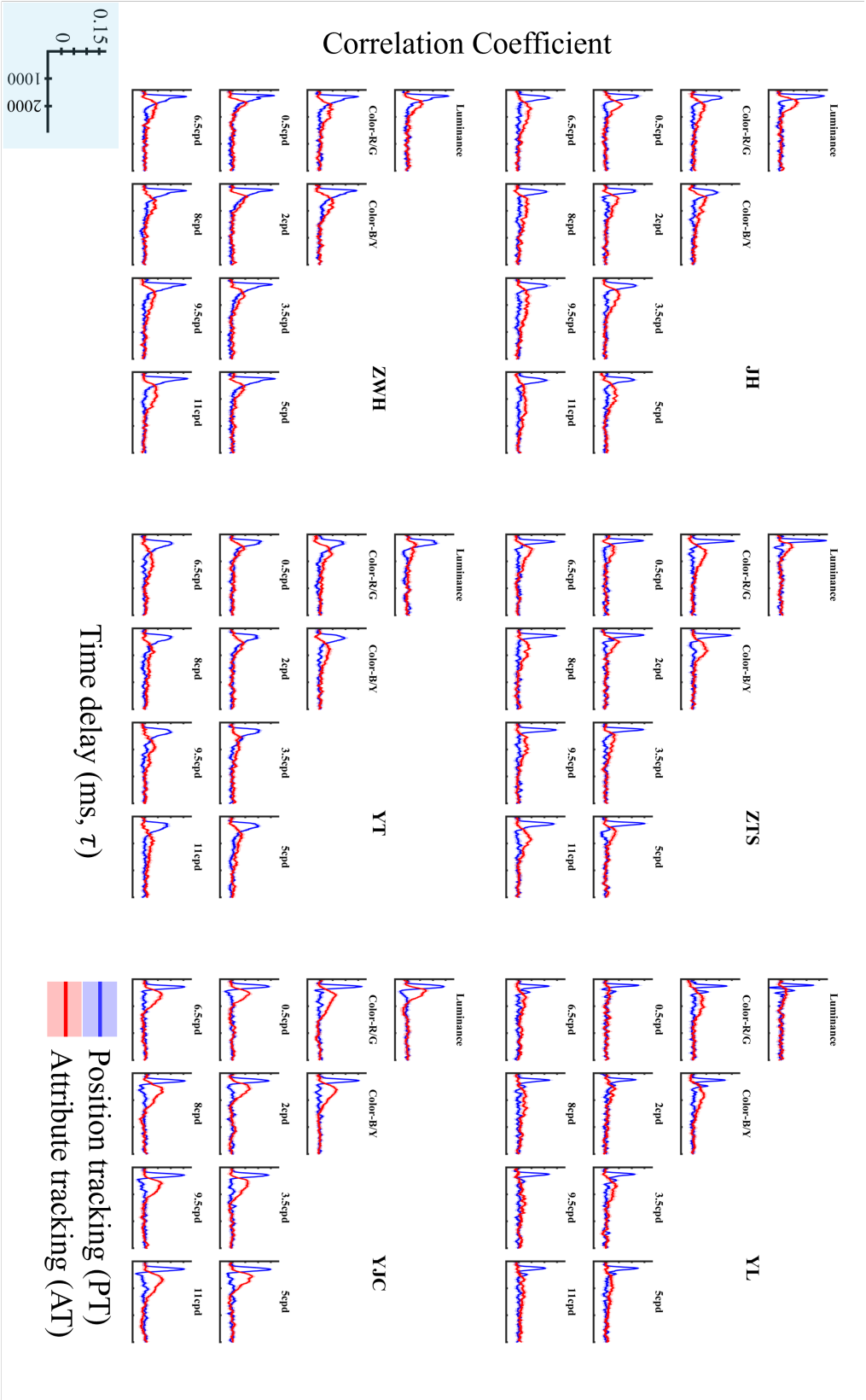


Figure 9-11. Individual CCG function for the Figure 6-2

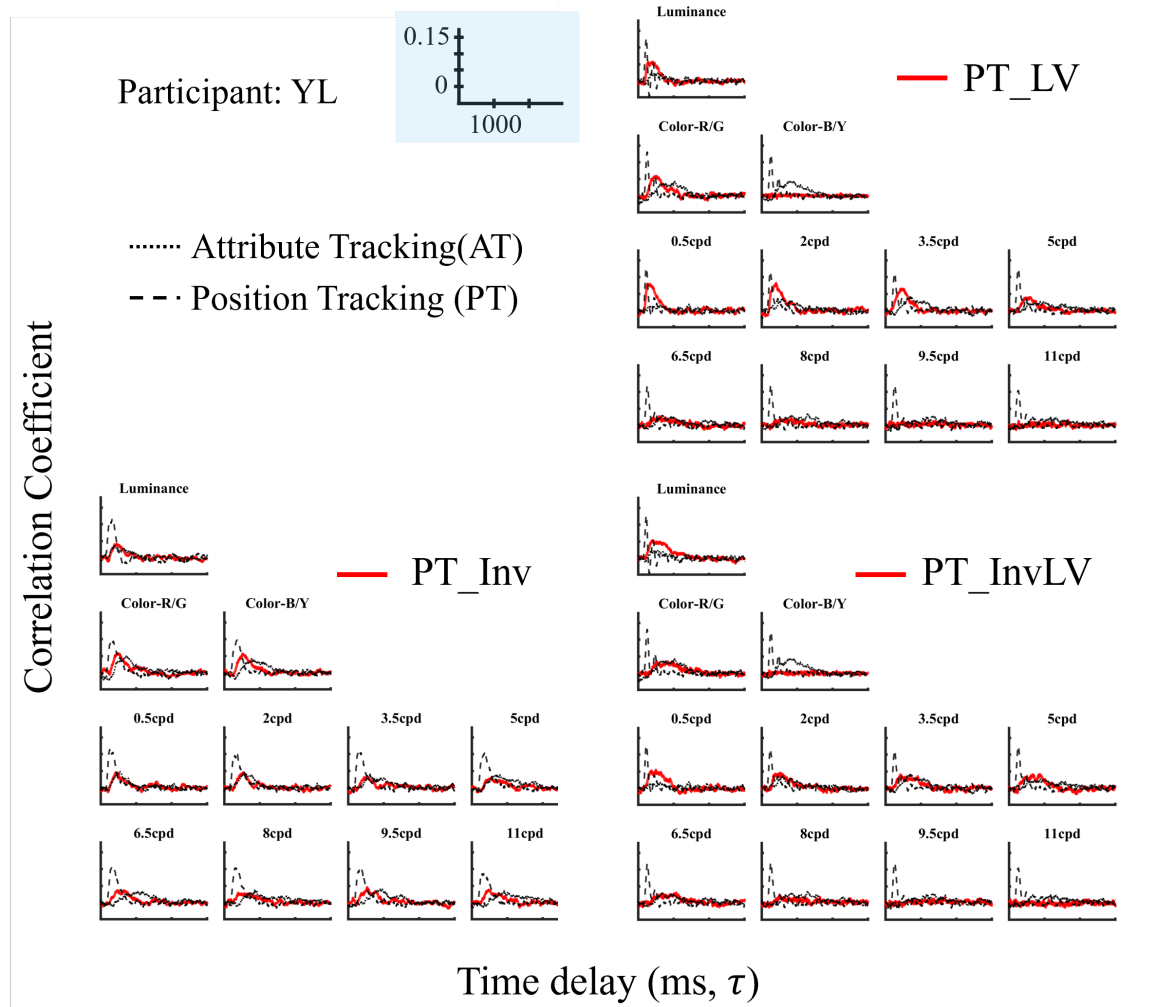


Figure 9-12. Individual YL's CCG function corresponds to the Figure 6-5

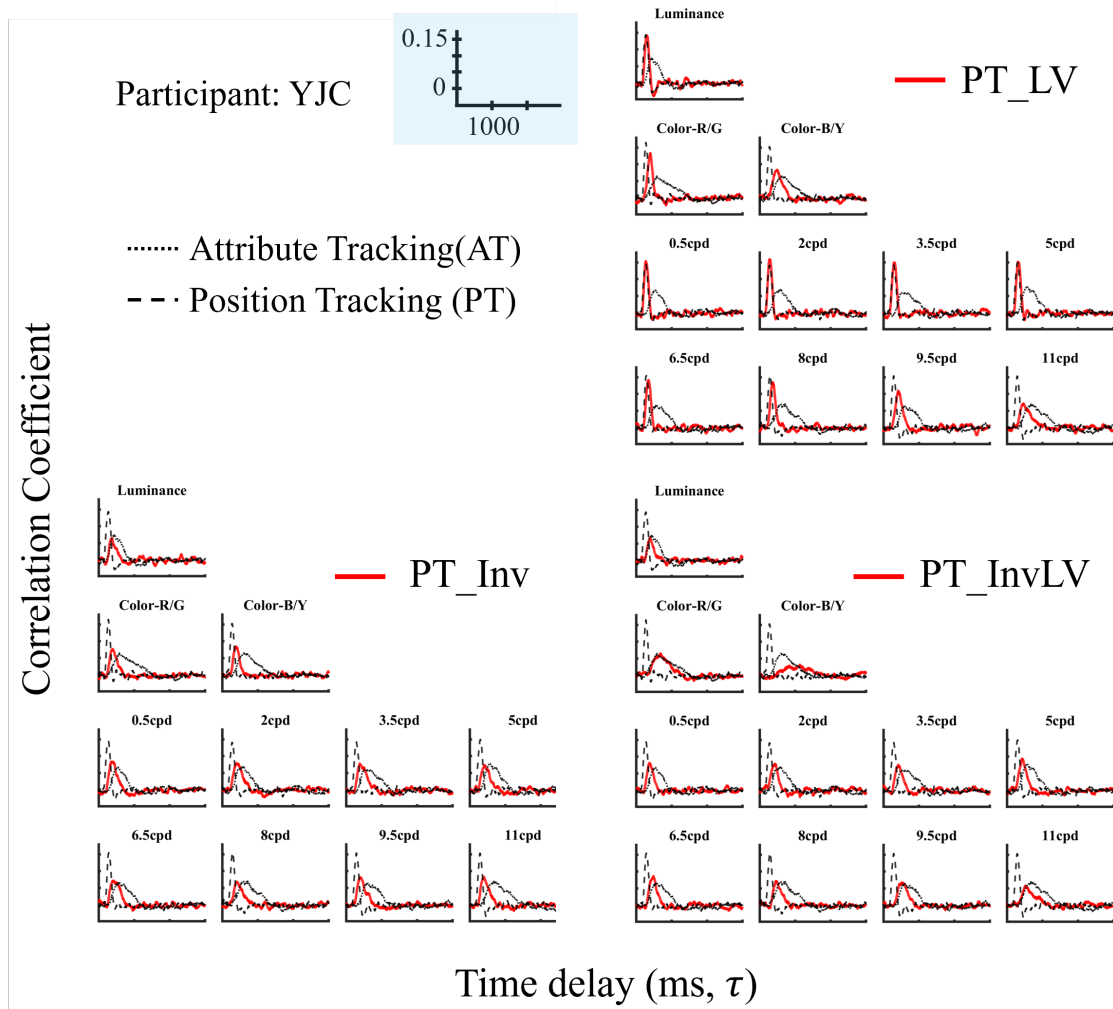


Figure 9-13. Individual YJC's CCG function corresponds to the Figure 6-5

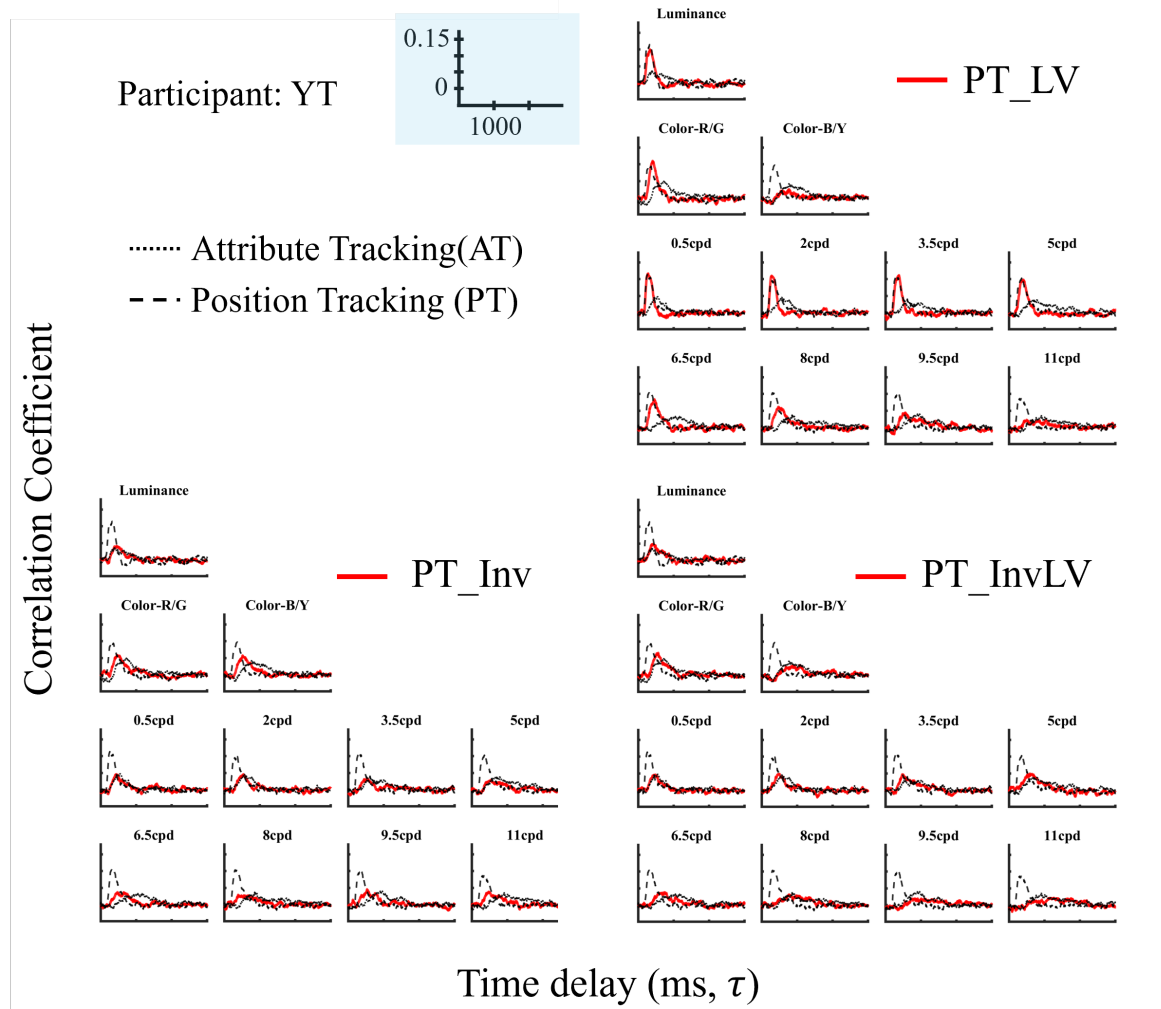


Figure 9-14. Individual YT's CCG function corresponds to the Figure 6-5

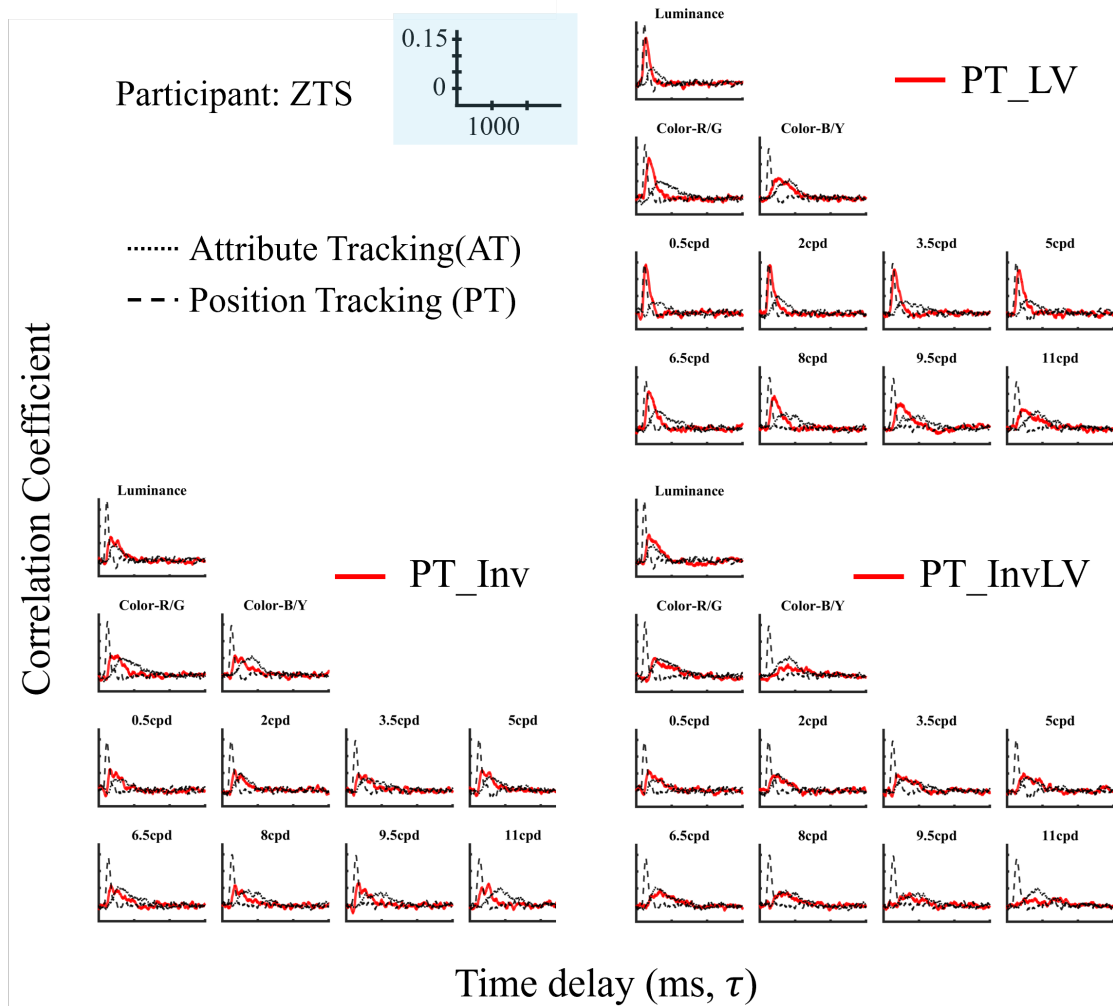


Figure 9-15. Individual ZTS's CCG function corresponds to the Figure 6-5

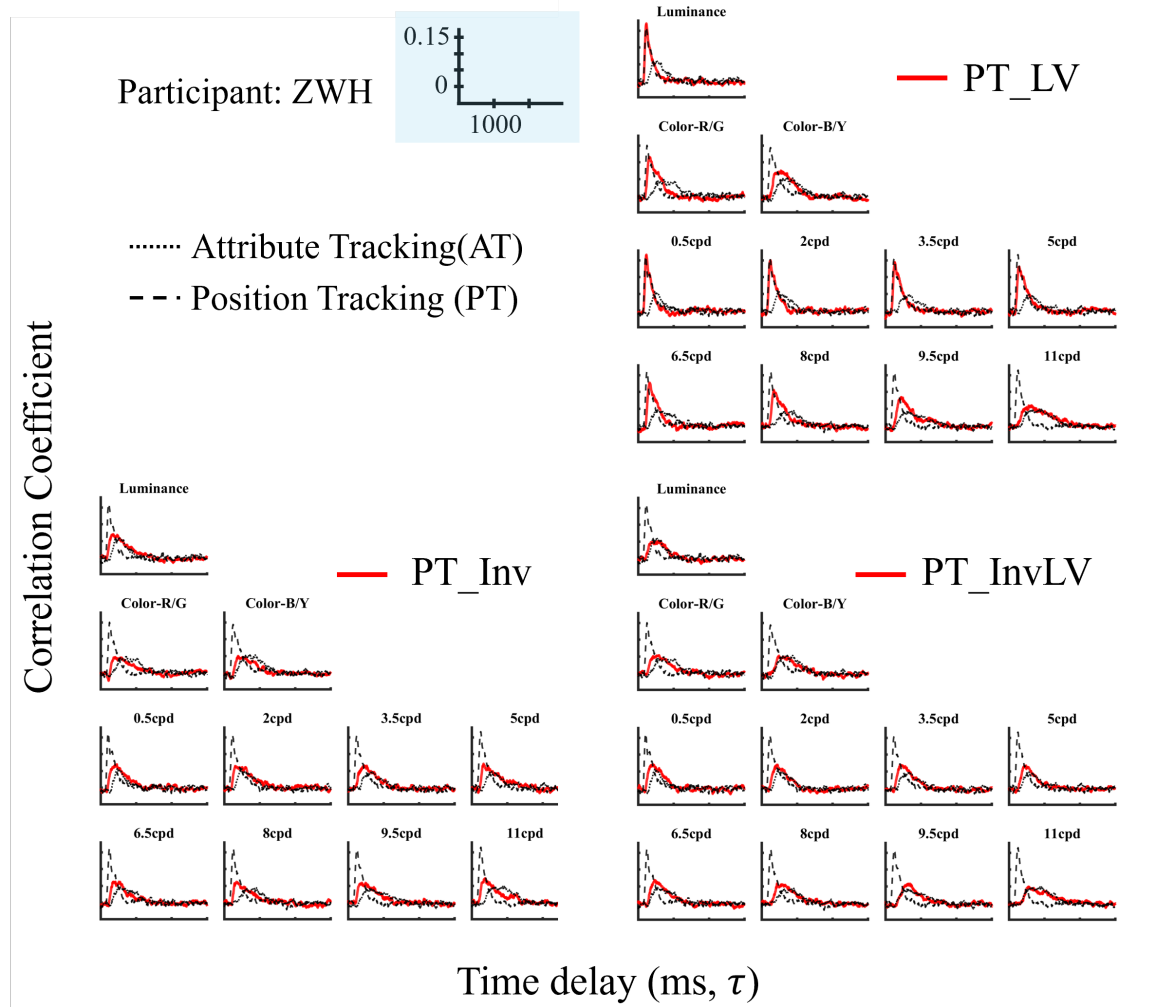


Figure 9-16. Individual ZWH's CCG function corresponds to the Figure 6-5

Reference

- Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, 2(2), 284. <https://doi.org/10.1364/JOSAA.2.000284>
- Adelson, E. H., & Farid, H. (1999). Filtering Reveals Form in Temporally Structured Displays. *Science*, 286(5448), 2231–2231. <https://doi.org/10.1126/science.286.5448.2231a>
- Adrian, E. D. (1946). Rod and cone components in the electric response of the eye. *The Journal of Physiology*, 105(1), 24–37.
- Alais, D., Blake, R., & Lee, S.-H. (1998). Visual features that vary together over time group together over space. *Nature Neuroscience*, 1(2), 160–164. <https://doi.org/10.1038/414>
- Ambrosi, P., Burr, D. C., & Cicchini, G. M. (2022). Ideal observer analysis for continuous tracking experiments. *Journal of Vision*, 22(2), 3. <https://doi.org/10.1167/jov.22.2.3>
- Ambrosi, P., Burr, D. C., & Morrone, M. C. (2024). Investigating cross-orientation inhibition with continuous tracking. *Journal of Vision*, 24(2), 2. <https://doi.org/10.1167/jov.24.2.2>
- Andersen, R. A., & Bradley, D. C. (1998). Perception of three-dimensional structure from motion. *Trends in Cognitive Sciences*, 2(6), 222–228. [https://doi.org/10.1016/S1364-6613\(98\)01181-4](https://doi.org/10.1016/S1364-6613(98)01181-4)
- Anderson, B. L. (2020). Mid-level vision. *Current Biology*, 30(3), R105–R109. <https://doi.org/10.1016/j.cub.2019.11.088>

- Anstis, S., & Cavanagh, P. (1983). *A minimum motion technique for judging equiluminance*. York University Toronto.
[http://wexler.free.fr/library/files/anstis%20\(1983\)%20a%20minimum%20motion%20technique%20for%20judging%20equiluminance.pdf](http://wexler.free.fr/library/files/anstis%20(1983)%20a%20minimum%20motion%20technique%20for%20judging%20equiluminance.pdf)
- Arsenault, E., Yoonessi, A., & Baker, C., Jr. (2011). Higher order texture statistics impair contrast boundary segmentation. *Journal of Vision*, 11(10), 14.
<https://doi.org/10.1167/11.10.14>
- Ashraf, M., Mantiuk, R. K., Chapiro, A., & Wuerger, S. (2024). castleCSF — A contrast sensitivity function of color, area, spatiotemporal frequency, luminance and eccentricity. *Journal of Vision*, 24(4), 5. <https://doi.org/10.1167/jov.24.4.5>
- Barnett, M. A., Chin, B. M., Aguirre, G. K., Burge, J., & Brainard, D. H. (2024). *Temporal dynamics of color processing measured using a continuous tracking task*. <https://doi.org/10.1101/2024.03.01.582975>
- Barnett, M. A., Chin, B. M., Aguirre, G. K., Burge, J., & Brainard, D. H. (2025). Temporal dynamics of human color processing measured using a continuous tracking task. *Journal of Vision*, 25(2), 12. <https://doi.org/10.1167/jov.25.2.12>
- Battelli, L., Cavanagh, P., Martini, P., & Barton, J. J. S. (2003). Bilateral deficits of transient visual attention in right parietal patients. *Brain*, 126(10), 2164–2174.
<https://doi.org/10.1093/brain/awg221>
- Battelli, L., Pascual-Leone, A., & Cavanagh, P. (2007). The ‘when’ pathway of the right parietal lobe. *Trends in Cognitive Sciences*, 11(5), 204–210.
<https://doi.org/10.1016/j.tics.2007.03.001>
- Beck, J. (1966). Effect of orientation and of shape similarity on perceptual grouping. *Perception & Psychophysics*, 1(5), 300–302.
<https://doi.org/10.3758/BF03207395>

- Bergen, J. R., & Adelson, E. H. (1988). Early vision and texture perception. *Nature*, 333(6171), 363–364. <https://doi.org/10.1038/333363a0>
- Bergen, James R., & Landy, Michael S. (1991). Computational Modeling of Visual Texture Segregation. In M. Landy & J. A. Movshon (Eds.), *Computational Models of Visual Processing*. The MIT Press. <https://doi.org/10.7551/mitpress/2002.003.0025>
- Blake, R., & Lee, S.-H. (2000). Temporal Structure in the Input to Vision Can Promote Spatial Grouping. In S.-W. Lee, H. H. Bülthoff, & T. Poggio (Eds.), *Biologically Motivated Computer Vision* (Vol. 1811, pp. 635–653). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45482-9_64
- Blake, R., & Lee, S.-H. (2005). The Role of Temporal Structure in Human Vision. *Behavioral and Cognitive Neuroscience Reviews*, 4(1), 21–42. <https://doi.org/10.1177/1534582305276839>
- Bonnen, K., Burge, J., Yates, J., Pillow, J., & Cormack, L. K. (2015). Continuous psychophysics: Target-tracking to measure visual sensitivity. *Journal of Vision*, 15(3), 14. <https://doi.org/10.1167/15.3.14>
- Bonnen, K., Huk, A. C., & Cormack, L. K. (2017). Dynamic mechanisms of visually guided 3D motion tracking. *Journal of Neurophysiology*, 118(3), 1515–1531. <https://doi.org/10.1152/jn.00831.2016>
- Born, R. T., & Bradley, D. C. (2005). STRUCTURE AND FUNCTION OF VISUAL AREA MT. *Annual Review of Neuroscience*, 28(Volume 28, 2005), 157–189. <https://doi.org/10.1146/annurev.neuro.26.041002.131052>
- Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10(4), 433–436. <https://doi.org/10.1163/156856897X00357>

- Breitmeyer, B. G., & Ganz, L. (1977). Temporal studies with flashed gratings: Inferences about human transient and sustained channels. *Vision Research*, 17(7), 861–865. [https://doi.org/10.1016/0042-6989\(77\)90130-4](https://doi.org/10.1016/0042-6989(77)90130-4)
- Brueggemann, J. (2007). The hand is NOT quicker than the eye. *Journal of Vision*, 7(15), 54. <https://doi.org/10.1167/7.15.54>
- Burge, J., & Cormack, L. K. (2020). *Target tracking reveals the time course of visual processing with millisecond-scale precision* [Preprint]. Neuroscience. <https://doi.org/10.1101/2020.08.05.238642>
- Burge, J., & Cormack, L. K. (2024). Continuous psychophysics shows millisecond-scale visual processing delays are faithfully preserved in movement dynamics. *Journal of Vision*, 24(5), 4. <https://doi.org/10.1167/jov.24.5.4>
- Burr, D. C., & Morrone, M. C. (1993). Impulse-response functions for chromatic and achromatic stimuli. *Journal of the Optical Society of America A*, 10(8), 1706. <https://doi.org/10.1364/JOSAA.10.001706>
- Bushara, K. O., Grafman, J., & Hallett, M. (2001). Neural Correlates of Auditory–Visual Stimulus Onset Asynchrony Detection. *Journal of Neuroscience*, 21(1), 300–304. <https://doi.org/10.1523/JNEUROSCI.21-01-00300.2001>
- Bushnell, B. N., Harding, P. J., Kosai, Y., & Pasupathy, A. (2011). Partial Occlusion Modulates Contour-Based Shape Encoding in Primate Area V4. *Journal of Neuroscience*, 31(11), 4012–4024. <https://doi.org/10.1523/JNEUROSCI.4766-10.2011>
- Callaghan, T. C., Lasaga, M. I., & Garner, W. R. (1986). Visual texture segregation based on orientation and hue. *Perception & Psychophysics*, 39(1), 32–38. <https://doi.org/10.3758/BF03207581>

- Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1), 51–62.
<https://doi.org/10.1038/nrn3136>
- Cavanagh, P., MacLeod, D. I. A., & Anstis, S. M. (1987). Equiluminance: Spatial and temporal factors and the contribution of blue-sensitive cones. *Journal of the Optical Society of America A*, 4(8), 1428.
<https://doi.org/10.1364/JOSAA.4.001428>
- Chen, Y.-J., Sun, Z., & Nishida, S. (2024a). Feature-invariant processing of spatial segregation based on temporal asynchrony. *Journal of Vision*, 24(5), 15.
<https://doi.org/10.1167/jov.24.5.15>
- Chen, Y.-J., Sun, Z., & Nishida, S. (2024b). Temporal Dynamics Gap between Position Tracking and Attribute Tracking. *Journal of Vision*, 24(10), 465.
<https://doi.org/10.1167/jov.24.10.465>
- Chen, Y.-J., Sun, Z., & Nishida, S. (2025). *Human visual grouping based on within- and cross-area temporal correlations*.
<https://doi.org/10.1101/2025.03.27.645664>
- Cook, P. B., & McReynolds, J. S. (1998). Lateral inhibition in the inner retina is important for spatial tuning of ganglion cells. *Nature Neuroscience*, 1(8), 714–719.
<https://doi.org/10.1038/3714>
- Craft, E., Schütze, H., Niebur, E., & von der Heydt, R. (2007). A Neural Model of Figure–Ground Organization. *Journal of Neurophysiology*, 97(6), 4310–4326.
<https://doi.org/10.1152/jn.00203.2007>
- Cropper, S. J. (2005). The detection of motion in chromatic stimuli: First-order and second-order spatial structure. *Vision Research*, 45(7), 865–880.
<https://doi.org/10.1016/j.visres.2004.09.043>

- Cropper, S. J., & Derrington, A. M. (1994). Motion of chromatic stimuli: First-order or second-order? *Vision Research*, 34(1), 49–58. [https://doi.org/10.1016/0042-6989\(94\)90256-9](https://doi.org/10.1016/0042-6989(94)90256-9)
- Dakin, S. (2001). Information limit on the spatial integration of local orientation signals. *Journal of the Optical Society of America A*, 18(5), 1016. <https://doi.org/10.1364/JOSAA.18.001016>
- Dakin, S. (2015). Seeing statistical regularities. In J. Wagemans (Ed.), *The Oxford Handbook of Perceptual Organization* (p. 0). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199686858.013.054>
- Dakin, S., & Watt, R. J. (1997). The computation of orientation statistics from visual texture. *Vision Research*, 37(22), 3181–3192. [https://doi.org/10.1016/S0042-6989\(97\)00133-8](https://doi.org/10.1016/S0042-6989(97)00133-8)
- DiMattina, C., & Baker, C. L. (2021). Segmenting surface boundaries using luminance cues. *Scientific Reports*, 11(1), 10074. <https://doi.org/10.1038/s41598-021-89277-2>
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* (arXiv:2010.11929). arXiv. <http://arxiv.org/abs/2010.11929>
- Driver, J., Davis, G., Russell, C., Turatto, M., & Freeman, E. (2001). Segmentation, attention and phenomenal visual objects. *Cognition*, 80(1–2), 61–95. [https://doi.org/10.1016/S0010-0277\(00\)00151-7](https://doi.org/10.1016/S0010-0277(00)00151-7)
- D’Souza, D. V., Auer, T., Strasburger, H., Frahm, J., & Lee, B. B. (2011). Temporal frequency and chromatic processing in humans: An fMRI study of the cortical visual areas. *Journal of Vision*, 11(8), 8. <https://doi.org/10.1167/11.8.8>

- Egley, R., Driver, J., & Rafal, R. D. (1994). Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General*, 123(2), 161–177. <https://doi.org/10.1037/0096-3445.123.2.161>
- Ejima, Y., & Ohtani, Y. (1987). Simple reaction time to sinusoidal grating and perceptual integration time: Contributions of perceptual and response processes. *Vision Research*, 27(2), 269–276. [https://doi.org/10.1016/0042-6989\(87\)90189-1](https://doi.org/10.1016/0042-6989(87)90189-1)
- Ejima, Y., & Ohtani, Y. (1989). Analysis of simple reaction time to sinusoidal grating by means of a linear filter model of the detection process. *Perception & Psychophysics*, 46(2), 119–126. <https://doi.org/10.3758/BF03204971>
- Elias, E., & Sweeny, T. D. (2020). Integration and segmentation conflict during ensemble coding of shape. *Journal of Experimental Psychology: Human Perception and Performance*, 46(6), 593–609. <https://doi.org/10.1037/xhp0000733>
- Ellis, S. R., & Adelstein, B. D. (2014). Interaction of Movement Difficulty due to the Misalignment Effect and Response Latency: Connections to Fitts' Law and Implications for System Latency Requirements. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 58(1), 1756–1760. <https://doi.org/10.1177/1541931214581366>
- Ellis, S. R., Adelstein, B. D., & Yeom, K. (2013). Misalignment Effect Function Measurement for Oblique Rotation Axes: Counterintuitive Predictions and Theoretical Extensions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 57(1), 1159–1163. <https://doi.org/10.1177/1541931213571258>

- El-Shamayleh, Y., & Movshon, J. A. (2011). Neuronal Responses to Texture-Defined Form in Macaque Visual Area V2. *Journal of Neuroscience*, 31(23), 8543–8555. <https://doi.org/10.1523/JNEUROSCI.5974-10.2011>
- Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, 392(6676), 598–601. <https://doi.org/10.1038/33402>
- Fahle, M. (1993). Figure–ground discrimination from temporal information. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 254(1341), 199–203. <https://doi.org/10.1098/rspb.1993.0146>
- Fahle, M., & Koch, C. (1995). Spatial displacement, but not temporal asynchrony, destroys figural binding. *Vision Research*, 35(4), 491–494. [https://doi.org/10.1016/0042-6989\(94\)00126-7](https://doi.org/10.1016/0042-6989(94)00126-7)
- Farid, H. (2002). Temporal synchrony in perceptual grouping: A critique. *Trends in Cognitive Sciences*, 6(7), 284–288. [https://doi.org/10.1016/S1364-6613\(02\)01927-7](https://doi.org/10.1016/S1364-6613(02)01927-7)
- Farid, H., & Adelson, E. H. (2001). Synchrony does not promote grouping in temporally structured displays. *Nature Neuroscience*, 4(9), 875–876. <https://doi.org/10.1038/nn0901-875>
- Fiedler, M. (1989). Laplacian of graphs and algebraic connectivity. *Banach Center Publications*, 25(1), 57–70. <https://doi.org/10.4064/-25-1-57-70>
- Field, D. J., Golden, J. R., & Hayes, A. (2013). *Contour integration and the association field*. MIT Press. <https://dr.ntu.edu.sg/handle/10356/105661>
- Field, D. J., Hayes, A., & Hess, R. F. (1993). Contour integration by the human visual system: Evidence for a local “association field.” *Vision Research*, 33(2), 173–193. [https://doi.org/10.1016/0042-6989\(93\)90156-Q](https://doi.org/10.1016/0042-6989(93)90156-Q)

- Folks, J. L., & Chhikara, R. S. (1978). The Inverse Gaussian Distribution and its Statistical Application—A Review. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(3), 263–275. <https://doi.org/10.1111/j.2517-6161.1978.tb01039.x>
- Forte, J., Hogben, J. H., & Ross, J. (1999). Spatial limitations of temporal segmentation. *Vision Research*, 39(24), 4052–4061. [https://doi.org/10.1016/S0042-6989\(99\)00124-8](https://doi.org/10.1016/S0042-6989(99)00124-8)
- Fredericksen, R. E., & Hess, R. F. (1997). Temporal detection in human vision: Dependence on stimulus energy. *Journal of the Optical Society of America A*, 14(10), 2557. <https://doi.org/10.1364/JOSAA.14.002557>
- Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience*, 16(7), 974–981. <https://doi.org/10.1038/nn.3402>
- Fujisaki, W., Kitazawa, S., & Nishida, S. (2012). Multisensory Timing. In B. E. Stein (Ed.), *The New Handbook of Multisensory Processing* (pp. 301–318). The MIT Press. <https://doi.org/10.7551/mitpress/8466.003.0026>
- Fujisaki, W., & Nishida, S. (2007). Feature-based processing of audio-visual synchrony perception revealed by random pulse trains. *Vision Research*, 47(8), 1075–1093. <https://doi.org/10.1016/j.visres.2007.01.021>
- Fujisaki, W., & Nishida, S. (2010). A common perceptual temporal limit of binding synchronous inputs across different sensory attributes and modalities. *Proceedings of the Royal Society B: Biological Sciences*, 277(1692), 2281–2290. <https://doi.org/10.1098/rspb.2010.0243>

- Garcia-Suarez, L., & Mullen, K. T. (2013). Form and motion processing of second-order stimuli in color vision. *Journal of Vision*, 13(7), 10–10. <https://doi.org/10.1167/13.7.10>
- Gaveau, V., Pisella, L., Priot, A.-E., Fukui, T., Rossetti, Y., Pélisson, D., & Prablanc, C. (2014). Automatic online control of motor adjustments in reaching and grasping. *Neuropsychologia*, 55, 25–40. <https://doi.org/10.1016/j.neuropsychologia.2013.12.005>
- Ghosh, A., Rothwell, J., & Haggard, P. (2014). Using voluntary motor commands to inhibit involuntary arm movements. *Proceedings of the Royal Society B: Biological Sciences*, 281(1794), 20141139. <https://doi.org/10.1098/rspb.2014.1139>
- Gilchrist, A. (2022). Gestalt Theory and the Network of Traditional Hypotheses. *Gestalt Theory*, 44, 97–116. <https://doi.org/10.2478/gth-2022-0003>
- Glass, L. (1969). Moiré Effect from Random Dots. *Nature*, 223(5206), 578–580. <https://doi.org/10.1038/223578a0>
- Glass, L., & Pérez, R. (1973). Perception of Random Dot Interference Patterns. *Nature*, 246(5432), Article 5432. <https://doi.org/10.1038/246360a0>
- Gomi, H. (2008). Implicit online corrections of reaching movements. *Current Opinion in Neurobiology*, 18(6), 558–564. <https://doi.org/10.1016/j.conb.2008.11.002>
- Graham, N., & Sutter, A. (2000). Normalization: Contrast-gain control in simple (Fourier) and complex (non-Fourier) pathways of pattern vision. *Vision Research*, 40(20), 2737–2761. [https://doi.org/10.1016/S0042-6989\(00\)00123-1](https://doi.org/10.1016/S0042-6989(00)00123-1)
- Groen, I. I. A., Piantoni, G., Montenegro, S., Flinker, A., Devore, S., Devinsky, O., Doyle, W., Dugan, P., Friedman, D., Ramsey, N. F., Petridou, N., & Winawer, J. (2022). Temporal Dynamics of Neural Responses in Human Visual Cortex. *The*

Journal of Neuroscience, 42(40), 7562–7580.

<https://doi.org/10.1523/JNEUROSCI.1812-21.2022>

Grossberg, S., & Pinna, B. (n.d.). Neural Dynamics of Gestalt Principles of Perceptual Organization: From Grouping to Shape and Meaning. *GESTALT THEORY*, 34.

Grunewald, A., Bradley, D. C., & Andersen, R. A. (2002). Neural Correlates of Structure-from-Motion Perception in Macaque V1 and MT. *The Journal of Neuroscience*, 22(14), 6195–6207. <https://doi.org/10.1523/JNEUROSCI.22-14-06195.2002>

Guttman, S. E., Gilroy, L. A., & Blake, R. (2005). Mixed messengers, unified message: Spatial grouping from temporal structure. *Vision Research*, 45(8), 1021–1030. <https://doi.org/10.1016/j.visres.2004.10.014>

Guttman, S. E., Gilroy, L. A., & Blake, R. (2007). Spatial grouping in human vision: Temporal structure trumps temporal synchrony. *Vision Research*, 47(2), 219–230. <https://doi.org/10.1016/j.visres.2006.09.012>

Haberman, J., & Whitney, D. (2011). Efficient summary statistical representation when change localization fails. *Psychonomic Bulletin & Review*, 18(5), 855–859. <https://doi.org/10.3758/s13423-011-0125-6>

Hamilton, R., Bees, M. A., Chaplin, C. A., & McCulloch, D. L. (2007). The luminance–response function of the human photopic electroretinogram: A mathematical model. *Vision Research*, 47(23), 2968–2972. <https://doi.org/10.1016/j.visres.2007.04.020>

Han, T., & Proctor, R. W. (2022). Revisiting variable-foreperiod effects: Evaluating the repetition priming account. *Attention, Perception & Psychophysics*, 84(4), 1193–1207. <https://doi.org/10.3758/s13414-022-02476-5>

- Harris, M. G., & Georgeson, M. A. (1986). Sustained and transient temporal integration functions depend on spatial frequency, not grating area. *Vision Research*, 26(11), 1779–1782. [https://doi.org/10.1016/0042-6989\(86\)90128-8](https://doi.org/10.1016/0042-6989(86)90128-8)
- He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.1512.03385>
- Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9(2), 181–197. <https://doi.org/10.1017/S0952523800009640>
- Heitger, F., von der Heydt, R., Peterhans, E., Rosenthaler, L., & Kübler, O. (1998). Simulation of neural contour mechanisms: Representing anomalous contours. *Image and Vision Computing*, 16(6), 407–421. [https://doi.org/10.1016/S0262-8856\(97\)00083-8](https://doi.org/10.1016/S0262-8856(97)00083-8)
- Helmholtz, H. (1872). On the Recent Advances in the Theory of Vision. *Buffalo Medical and Surgical Journal*, 11(7), 245–256.
- Holcombe, A. O. (2009). Seeing slow and seeing fast: Two limits on perception. *Trends in Cognitive Sciences*, 13(5), 216–221. <https://doi.org/10.1016/j.tics.2009.02.005>
- Horiguchi, H., Nakadomari, S., Misaki, M., & Wandell, B. A. (2009). Two temporal channels in human V1 identified using fMRI. *NeuroImage*, 47(1), 273–280. <https://doi.org/10.1016/j.neuroimage.2009.03.078>
- Houshmand Chatroudi, A., Mioni, G., & Yotsumoto, Y. (2024). On the nonlinearity of the foreperiod effect. *Scientific Reports*, 14(1), 2780. <https://doi.org/10.1038/s41598-024-53347-y>
- Hu, L., Wang, C., & Zhang, X. (2022). Spatial resolution and object segmentation efficiency constrain grouping effects in attentive tracking. *Current Psychology*, 41(11), 7574–7587. <https://doi.org/10.1007/s12144-020-01277-z>

- Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1), 106-154.2.
- Hummel, J. E., & Biederman, I. (1992). *Dynamic Binding in a Neural Network for Shape Recognition*. <https://doi.org/10.1037/0033-295x.99.3.480>
- Hunt, C., & Meinhardt, G. (2021). Synergy of spatial frequency and orientation bandwidth in texture segregation. *Journal of Vision*, 21(2), 5. <https://doi.org/10.1167/jov.21.2.5>
- Irvin, G. E., Casagrande, V. A., & Norton, T. T. (1993). Center/surround relationships of magnocellular, parvocellular, and koniocellular relay cells in primate lateral geniculate nucleus. *Visual Neuroscience*, 10(2), 363–373. <https://doi.org/10.1017/S0952523800003758>
- Jäkel, F., Singh, M., Wichmann, F. A., & Herzog, M. H. (2016). An overview of quantitative approaches in Gestalt perception. *Vision Research*, 126, 3–8. <https://doi.org/10.1016/j.visres.2016.06.004>
- Jeurissen, D., Self, M. W., & Roelfsema, P. R. (2016). Serial grouping of 2D-image regions with object-based attention in humans. *eLife*, 5, e14320. <https://doi.org/10.7554/eLife.14320>
- Jianbo Shi, & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 888–905. <https://doi.org/10.1109/34.868688>
- Jiang, P., Ergu, D., Liu, F., Cai, Y., & Ma, B. (2022). A Review of Yolo Algorithm Developments. *Procedia Computer Science*, 199, 1066–1073. <https://doi.org/10.1016/j.procs.2022.01.135>

- Jiang, Y., Zhou, K., & He, S. (2007). Human Visual Cortex Responds to Invisible Chromatic Flicker. *Nature Neuroscience*, 10, 657–662. <https://doi.org/10.1038/nn1879>
- Jolicoeur, P., Ullman, S., & Mackay, M. (1986). Curve tracing: A possible basic operation in the perception of spatial relations. *Memory & Cognition*, 14(2), 129–140. <https://doi.org/10.3758/BF03198373>
- Julesz, B. (1962). Visual Pattern Discrimination. *IRE Transactions on Information Theory*, 8(2), 84–92. <https://doi.org/10.1109/TIT.1962.1057698>
- Julesz, B. (1965). Texture and Visual Perception. *Scientific American*, 212(2), 38–49.
- Kandil, F. I., & Fahle, M. (2001). Purely temporal figure–ground segregation. *European Journal of Neuroscience*, 13(10), 2004–2008.
- Kandil, F. I., & Fahle, M. (2004). Figure–ground segregation can rely on differences in motion direction. *Vision Research*, 44(27), 3177–3182. <https://doi.org/10.1016/j.visres.2004.07.027>
- Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *Journal of Neuroscience*, 17(11), 4302–4311. <https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997>
- Kaplan, E. (2003). The M, P, and K Pathways of the Primate Visual System. In L. M. Chalupa & J. S. Werner (Eds.), *The Visual Neurosciences*, 2-vol. Set (pp. 481–493). The MIT Press. <https://doi.org/10.7551/mitpress/7131.003.0036>
- Karlin, L. (1959). Reaction time as a function of foreperiod duration and variability. *Journal of Experimental Psychology*, 58(2), 185–191. <https://doi.org/10.1037/h0049152>

- Kellman, P. J., & Shipley, T. F. (1991). A theory of visual interpolation in object perception. *Cognitive Psychology*, 23(2), 141–221. [https://doi.org/10.1016/0010-0285\(91\)90009-D](https://doi.org/10.1016/0010-0285(91)90009-D)
- Kelly, D. H. (1979). Motion and vision II Stabilized spatio-temporal threshold surface. *Journal of the Optical Society of America*, 69(10), 1340. <https://doi.org/10.1364/JOSA.69.001340>
- Kenner, N. M., Mumford, J. A., Hommer, R. E., Skup, M., Leibenluft, E., & Poldrack, R. A. (2010). Inhibitory Motor Control in Response Stopping and Response Switching. *Journal of Neuroscience*, 30(25), 8512–8518. <https://doi.org/10.1523/JNEUROSCI.1096-10.2010>
- Kingdom, F. a. A., Prins, N., & Hayes, A. (2003). Mechanism independence for texture-modulation detection is consistent with a filter-rectify-filter mechanism. *Visual Neuroscience*, 20(1), 65–76. <https://doi.org/10.1017/S0952523803201073>
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). *Segment Anything* (arXiv:2304.02643). arXiv. <https://doi.org/10.48550/arXiv.2304.02643>
- Köhler, W. (1930). Some tasks of Gestalt psychology. In *Psychologies of 1930*. (pp. 143–160). Clark University Press. <https://doi.org/10.1037/11017-008>
- Kojima, H. (1998). Figure/ground segregation from temporal delay is best at high spatial frequencies. *Vision Research*, 38(23), 3729–3734. [https://doi.org/10.1016/S0042-6989\(98\)00012-1](https://doi.org/10.1016/S0042-6989(98)00012-1)
- Konen, C. S., & Kastner, S. (2008). Two hierarchically organized neural systems for object information in human visual cortex. *Nature Neuroscience*, 11(2), 224–231. <https://doi.org/10.1038/nn2036>

- Kosslyn, S. M., & Osherson, D. N. (Eds.). (1995). Visual Surface Representation. In *An Invitation to Cognitive Science*. The MIT Press. <https://doi.org/10.7551/mitpress/3965.003.0004>
- Kourtzi, Z., & Kanwisher, N. (2001). Representation of Perceived Object Shape by the Human Lateral Occipital Complex. *Science*, 293(5534), 1506–1509. <https://doi.org/10.1126/science.1061133>
- Krähenbühl, P., & Koltun, V. (2011). Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. *Advances in Neural Information Processing Systems*, 24. https://proceedings.neurips.cc/paper_files/paper/2011/hash/beda24c1e1b46055dff2c39c98fd6fc1-Abstract.html
- Kremers, J., Baraas, R. C., & Marshall, N. J. (Eds.). (2016). *Human Color Vision*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-44978-4>
- Kubilius, J., Wagemans, J., & Op de Beeck, H. (2014a). A conceptual framework of computations in mid-level vision. *Frontiers in Computational Neuroscience*, 8. <https://www.frontiersin.org/articles/10.3389/fncom.2014.00158>
- Kubilius, J., Wagemans, J., & Op de Beeck, H. P. (2014b). Encoding of configural regularity in the human visual system. *Journal of Vision*, 14(9), 11. <https://doi.org/10.1167/14.9.11>
- Kulikowski, J. J., Robson, A. G., & Murray, I. J. (2002). Scalp VEPs and intra-cortical responses to chromatic and achromatic stimuli in primates. *Documenta Ophthalmologica*, 105(2), 243–279. <https://doi.org/10.1023/A:1020557105243>
- Kulikowski, J. J., & Tolhurst, D. J. (1973). Psychophysical evidence for sustained and transient detectors in human vision. *The Journal of Physiology*, 232(1), 149–162. <https://doi.org/10.1113/jphysiol.1973.sp010261>

- Lamme, V. A. F., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, 23(11), 571–579. [https://doi.org/10.1016/S0166-2236\(00\)01657-X](https://doi.org/10.1016/S0166-2236(00)01657-X)
- Landis, C. (1954). Determinants of the Critical Flicker-Fusion Threshold. *Physiological Reviews*, 34(2), 259–286. <https://doi.org/10.1152/physrev.1954.34.2.259>
- Landy, M. S., & Kojima, H. (2001). Ideal cue combination for localizing texture-defined edges. *Journal of the Optical Society of America A*, 18(9), 2307. <https://doi.org/10.1364/JOSAA.18.002307>
- Landy, Michael S., & Graham, N. (2003). Visual Perception of Texture. In L. M. Chalupa & John S. Werner (Eds.), *The Visual Neurosciences*, 2-vol. Set (pp. 1106–1118). The MIT Press. <https://doi.org/10.7551/mitpress/7131.003.0084>
- Lappin, J., & Bell, H. (2021). Form and Function in Information for Visual Perception. *I-Perception*, 12, 204166952110533. <https://doi.org/10.1177/20416695211053352>
- Ledgeway, T., & Hutchinson, C. V. (2008). Choice reaction times for identifying the direction of first-order motion and different varieties of second-order motion. *Vision Research*, 48(2), 208–222. <https://doi.org/10.1016/j.visres.2007.11.008>
- Lee, S., Kim, K., Kim, J.-Y., Kim, M., & Yoo, H.-J. (2010). Familiarity based unified visual attention model for fast and robust object recognition. *Pattern Recognition*, 43(3), 1116–1128. <https://doi.org/10.1016/j.patcog.2009.07.014>
- Lee, S.-H., & Blake, R. (1999). Visual Form Created Solely from Temporal Structure. *Science*, 284(5417), 1165–1168. <https://doi.org/10.1126/science.284.5417.1165>
- Lee, S.-H., & Blake, R. (2001). Neural synergy in visual grouping: When good continuation meets common fate. *Vision Research*, 41(16), 2057–2064. [https://doi.org/10.1016/S0042-6989\(01\)00086-4](https://doi.org/10.1016/S0042-6989(01)00086-4)

- Legge, G. E. (1978). Sustained and transient mechanisms in human vision: Temporal and spatial properties. *Vision Research*, 18(1), 69–81.
[https://doi.org/10.1016/0042-6989\(78\)90079-2](https://doi.org/10.1016/0042-6989(78)90079-2)
- Leonards, U., & Singer, W. (1998). Two segmentation mechanisms with differential sensitivity for colour and luminance contrast. *Vision Research*, 38(1), 101–109.
[https://doi.org/10.1016/S0042-6989\(97\)00148-X](https://doi.org/10.1016/S0042-6989(97)00148-X)
- Leonards, U., Singer, W., & Fahle, M. (1996). The Influence of Temporal Phase Differences on Texture Segmentation. *Vision Research*, 36(17), 2689–2697.
[https://doi.org/10.1016/0042-6989\(96\)86829-5](https://doi.org/10.1016/0042-6989(96)86829-5)
- Li, Y., Wang, Y., & Li, S. (2019). Recurrent Processing of Contour Integration in the Human Visual Cortex as Revealed By fMRI-Guided TMS. *Cerebral Cortex*, 29(1), 17–26. <https://doi.org/10.1093/cercor/bhx296>
- Lit, A., & Hyman, A. (1951). THE MAGNITUDE OF THE PULFRICH STEREOPHENOMENON AS A FUNCTION OF DISTANCE OF OBSERVATION: *Optometry and Vision Science*, 28(11), 564–580.
<https://doi.org/10.1097/00006324-195111000-00004>
- Liu, T. (2019). Feature-based attention: Effects and control. *Current Opinion in Psychology*, 29, 187–192. <https://doi.org/10.1016/j.copsyc.2019.03.013>
- Malcolm, G. L., & Shomstein, S. (2015). Object-based attention in real-world scenes. *Journal of Experimental Psychology: General*, 144(2), 257–263.
<https://doi.org/10.1037/xge0000060>
- Malik, J., & Perona, P. (1990). Preattentive texture discrimination with early vision mechanisms. *Journal of the Optical Society of America A*, 7(5), 923.
<https://doi.org/10.1364/JOSAA.7.000923>

- Manahova, M. E., Spaak, E., & de Lange, F. P. (2020). Familiarity Increases Processing Speed in the Visual System. *Journal of Cognitive Neuroscience*, 32(4), 722–733. https://doi.org/10.1162/jocn_a_01507
- Mantiuk, R. K., Ashraf, M., & Chapiro, A. (2022). stelaCSF: A unified model of contrast sensitivity as the function of spatio-temporal frequency, eccentricity, luminance and area. *ACM Transactions on Graphics*, 41(4), 1–16. <https://doi.org/10.1145/3528223.3530115>
- Marmarelis, P. Z., & Marmarelis, V. Z. (1978). The White-Noise Method in System Identification. In P. Z. Marmarelis & V. Z. Marmarelis (Eds.), *Analysis of Physiological Systems: The White-Noise Approach* (pp. 131–180). Springer US. https://doi.org/10.1007/978-1-4613-3970-0_4
- Marr, D. (2010). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. MIT Press.
- Maruya, K., Holcombe, A. O., & Nishida, S. (2013). Rapid encoding of relationships between spatially remote motion signals. *Journal of Vision*, 13(2), 4–4. <https://doi.org/10.1167/13.2.4>
- Maunsell, J. H. R., Ghose, G. M., Assad, J. A., McADAMS, C. J., Boudreau, C. E., & Noerager, B. D. (1999). Visual response latencies of magnocellular and parvocellular LGN neurons in macaque monkeys. *Visual Neuroscience*, 16(1), 1–14. <https://doi.org/10.1017/S0952523899156177>
- Meese, T. S. (2010). Spatially extensive summation of contrast energy is revealed by contrast detection of micro-pattern textures. *Journal of Vision*, 10(8), 14. <https://doi.org/10.1167/10.8.14>
- Mehrani, P., & Tsotsos, J. K. (2023). *Self-attention in Vision Transformers Performs Perceptual Grouping, Not Attention*. <http://arxiv.org/abs/2303.01542>

- Mendola, J. D., Conner, I. P., Sharma, S., Bahekar, A., & Lemieux, S. (2006). fMRI Measures of Perceptual Filling-in in the Human Visual Cortex. *Journal of Cognitive Neuroscience*, 18(3), 363–375. <https://doi.org/10.1162/jocn.2006.18.3.363>
- Miyamoto, T., Miura, K., Kizuka, T., & Ono, S. (2020). Properties of smooth pursuit and visual motion reaction time to second-order motion stimuli. *PLOS ONE*, 15(12), e0243430. <https://doi.org/10.1371/journal.pone.0243430>
- Morgan, M., & Castet, E. (2002). High temporal frequency synchrony is insufficient for perceptual grouping. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 269(1490), 513–516. <https://doi.org/10.1098/rspb.2001.1920>
- Morrone, M. C., Denti, V., & Spinelli, D. (2002). Color and Luminance Contrasts Attract Independent Attention. *Current Biology*, 12(13), 1134–1137. [https://doi.org/10.1016/S0960-9822\(02\)00921-1](https://doi.org/10.1016/S0960-9822(02)00921-1)
- Motoyoshi, I. (2004). The role of spatial interactions in perceptual synchrony. *Journal of Vision*, 4(5), 1–1. <https://doi.org/10.1167/4.5.1>
- Motoyoshi, I., & Nishida, S. (2001). Visual response saturation to orientation contrast in the perception of texture boundary. *Journal of the Optical Society of America A*, 18(9), 2209. <https://doi.org/10.1364/JOSAA.18.002209>
- Motoyoshi, I., & Nishida, S. (2002). Spatiotemporal interactions in detection of texture orientation modulations. *Vision Research*, 42(26), 2829–2841. [https://doi.org/10.1016/S0042-6989\(02\)00336-X](https://doi.org/10.1016/S0042-6989(02)00336-X)
- Motoyoshi, I., & Nishida, S. (2004). Cross-orientation summation in texture segregation. *Vision Research*, 44(22), 2567–2576. <https://doi.org/10.1016/j.visres.2004.05.024>

- Moutoussis, K., & Zeki, S. (1997). A direct demonstration of perceptual asynchrony in vision. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 264(1380), 393–399. <https://doi.org/10.1098/rspb.1997.0056>
- Mullen, K. T. (2019). The response to colour in the human visual cortex: The fMRI approach. *Current Opinion in Behavioral Sciences*, 30, 141–148. <https://doi.org/10.1016/j.cobeha.2019.08.001>
- Mullen, K. T., Thompson, B., & Hess, R. F. (2010). Responses of the human visual cortex and LGN to achromatic and chromatic temporal modulations: An fMRI study. *Journal of Vision*, 10(13), 13. <https://doi.org/10.1167/10.13.13>
- Mulligan, J. B. (1997). Image processing for improved eye-tracking accuracy. *Behavior Research Methods, Instruments, & Computers*, 29(1), 54–65. <https://doi.org/10.3758/BF03200567>
- Mulligan, J., Stevenson, S., & Cormack, L. (2013). Reflexive and voluntary control of smooth eye movements. *Proceedings of SPIE - The International Society for Optical Engineering*, 8651. <https://doi.org/10.1117/12.2010333>
- Murray, S. O. (2003). Processing Shape, Motion and Three-dimensional Shape-from-motion in the Human Cortex. *Cerebral Cortex*, 13(5), 508–516. <https://doi.org/10.1093/cercor/13.5.508>
- Nakayama, K. (1971). Local adaptation in cat LGN cells: Evidence for a surround antagonism. *Vision Research*, 11(6), 501–509. [https://doi.org/10.1016/0042-6989\(71\)90073-3](https://doi.org/10.1016/0042-6989(71)90073-3)
- Nakayama, K., He, Z. J., & Shimojo, S. (1995). Visual surface representation: A critical link between lower-level and higher-level vision. In *Visual cognition: An invitation to cognitive science, Vol. 2, 2nd ed* (pp. 1–70). The MIT Press.

- Neumann, H., & Sepp, W. (1999). Recurrent V1–V2 interaction in early visual boundary processing. *Biological Cybernetics*, 81(5), 425–444. <https://doi.org/10.1007/s004220050573>
- Newsome, W. T., Shadlen, M. N., Zohary, E., Britten, K. H., & Movshon, J. A. (1995). Visual Motion: Linking Neuronal Activity to Psychophysical Performance. In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences*. MIT Press.
- Nishida, S., & Johnston, A. (2002). Marker Correspondence, Not Processing Latency, Determines Temporal Binding of Visual Attributes. *Current Biology*, 12(5), 359–368. [https://doi.org/10.1016/S0960-9822\(02\)00698-X](https://doi.org/10.1016/S0960-9822(02)00698-X)
- Nothdurft, H. C. (1985). Sensitivity for structure gradient in texture discrimination tasks. *Vision Research*, 25(12), 1957–1968. [https://doi.org/10.1016/0042-6989\(85\)90020-3](https://doi.org/10.1016/0042-6989(85)90020-3)
- Nothdurft, H.-C. (2007). Common Properties of Visual Segmentation. In *Ciba Foundation Symposium 184—Higher-Order Processing in the Visual System* (pp. 245–271). John Wiley & Sons, Ltd. <https://doi.org/10.1002/9780470514610.ch13>
- O'Donnell, B., & Colombo, E. (2008). Simple reaction times to chromatic stimuli: Luminance and chromatic contrast. *Lighting Research & Technology*, 40(4), 359–371. <https://doi.org/10.1177/1477153508097709>
- Oja, E. (1982). Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3), 267–273. <https://doi.org/10.1007/BF00275687>
- Okazawa, G., Sha, L., Purcell, B. A., & Kiani, R. (2018). Psychophysical reverse correlation reflects both sensory and decision-making processes. *Nature Communications*, 9, 3479. <https://doi.org/10.1038/s41467-018-05797-y>

- Op de Beeck, H., Wagemans, J., & Vogels, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nature Neuroscience*, 4(12), 1244–1252. <https://doi.org/10.1038/nn767>
- Orban, G. A., Dupont, P., De Bruyn, B., Vogels, R., Vandenberghe, R., & Mortelmans, L. (1995). A motion area in human visual cortex. *Proceedings of the National Academy of Sciences*, 92(4), 993–997. <https://doi.org/10.1073/pnas.92.4.993>
- Parise, C. V., & Ernst, M. O. (2016). Correlation detection as a general mechanism for multisensory integration. *Nature Communications*, 7(1), 11543. <https://doi.org/10.1038/ncomms11543>
- Parton, A., Donnelly, N., & Usher, M. (2001). The effects of temporal synchrony on the perceived organization of elements in spatially symmetric and asymmetric grids. *Visual Cognition*, 8(3–5), 637–654. <https://doi.org/10.1080/13506280143000179>
- Pasupathy, A., & Connor, C. E. (2001). Shape Representation in Area V4: Position-Specific Tuning for Boundary Conformation. *Journal of Neurophysiology*, 86(5), 2505–2519. <https://doi.org/10.1152/jn.2001.86.5.2505>
- Pasupathy, A., & Connor, C. E. (2002). Population coding of shape in area V4. *Nature Neuroscience*, 5(12), 1332–1338. <https://doi.org/10.1038/972>
- Peirce, J. W. (2015). Understanding mid-level representations in visual processing. *Journal of Vision*, 15(7), 5. <https://doi.org/10.1167/15.7.5>
- Pesnot-Lerousseau, J., Parise, C., Ernst, M., & Van Wassenhove, V. (2022). Multisensory correlation computations in the human brain identified by a time-resolved encoding model. *Nature Communications*, 13, 2489. <https://doi.org/10.1038/s41467-022-29687-6>

- Peterson, M. A., & Gibson, B. S. (1994). Must Figure-Ground Organization Precede Object Recognition? An Assumption in Peril. *Psychological Science*, 5(5), 253–259. <https://doi.org/10.1111/j.1467-9280.1994.tb00622.x>
- Pisella, L., Gréa, H., Tilikete, C., Vighetto, A., Desmurget, M., Rode, G., Boisson, D., & Rossetti, Y. (2000). An ‘automatic pilot’ for the hand in human posterior parietal cortex: Toward reinterpreting optic ataxia. *Nature Neuroscience*, 3(7), 729–736. <https://doi.org/10.1038/76694>
- Polat, U., & Sagi, D. (1993). Lateral interactions between spatial channels: Suppression and facilitation revealed by lateral masking experiments. *Vision Research*, 33(7), 993–999. [https://doi.org/10.1016/0042-6989\(93\)90081-7](https://doi.org/10.1016/0042-6989(93)90081-7)
- Poort, J., Raudies, F., Wannig, A., Lamme, V. A. F., Neumann, H., & Roelfsema, P. R. (2012). The Role of Attention in Figure-Ground Segregation in Areas V1 and V4 of the Visual Cortex. *Neuron*, 75(1), 143–156. <https://doi.org/10.1016/j.neuron.2012.04.032>
- Portilla, J., & Simoncelli, E. P. (2000). A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients. *International Journal of Computer Vision*, 40(1), 49–70. <https://doi.org/10.1023/A:1026553619983>
- Qiu, F. T., Sugihara, T., & von der Heydt, R. (2007). Figure-ground mechanisms provide structure for selective attention. *Nature Neuroscience*, 10(11), 1492–1499. <https://doi.org/10.1038/nn1989>
- Ramachandran, V. S., & Rogers-Ramachandran, D. C. (1991). Phantom contours: A new class of visual patterns that selectively activates the magnocellular pathway in man. *Bulletin of the Psychonomic Society*, 29(5), 391–394. <https://doi.org/10.3758/BF03333951>

- Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin & Review*, 9(2), 278–291. <https://doi.org/10.3758/BF03196283>
- Rensink, R. A. (2000). The Dynamic Representation of Scenes. *Visual Cognition*, 7(1–3), 17–42. <https://doi.org/10.1080/135062800394667>
- Rideaux, R., Badcock, D. R., Johnston, A., & Edwards, M. (2016). Temporal synchrony is an effective cue for grouping and segmentation in the absence of form cues. *Journal of Vision*, 16(11), 23. <https://doi.org/10.1167/16.11.23>
- Ringach, D., & Shapley, R. (2004). Reverse correlation in neurophysiology. *Cognitive Science*, 28(2), 147–166. https://doi.org/10.1207/s15516709cog2802_2
- Rodrigues, J., & du Buf, J. M. H. (2006). Multi-scale keypoints in V1 and beyond: Object segregation, scale selection, saliency maps and face detection. *Biosystems*, 86(1–3), 75–90. <https://doi.org/10.1016/j.biosystems.2006.02.019>
- Roelfsema, P. R. (2006). Cortical Algorithms for Perceptual Grouping. *Annual Review of Neuroscience*, 29(1), 203–227. <https://doi.org/10.1146/annurev.neuro.29.051605.112939>
- Roelfsema, P. R., Lamme, V. A. F., Spekreijse, H., & Bosch, H. (2002). Figure—Ground Segregation in a Recurrent Network Architecture. *Journal of Cognitive Neuroscience*, 14(4), 525–537. <https://doi.org/10.1162/08989290260045756>
- Rosenholtz, R. (2000). Significantly Different Textures: A Computational Model of Pre-attentive Texture Segmentation. In D. Vernon (Ed.), *Computer Vision—ECCV 2000* (pp. 197–211). Springer. https://doi.org/10.1007/3-540-45053-X_13

- Rucci, M., & Victor, J. D. (2015). The unsteady eye: An information-processing stage, not a bug. *Trends in Neurosciences*, 38(4), 195–206.
<https://doi.org/10.1016/j.tins.2015.01.005>
- Saarela, T. P., & Landy, M. S. (2012). Combination of texture and color cues in visual segmentation. *Vision Research*, 58, 59–67.
<https://doi.org/10.1016/j.visres.2012.01.019>
- Saarela, T. P., & Landy, M. S. (2015). Integration Trumps Selection in Object Recognition. *Current Biology*, 25(7), 920–927.
<https://doi.org/10.1016/j.cub.2015.01.068>
- Saijo, N., Murakami, I., Nishida, S., & Gomi, H. (2005). Large-Field Visual Motion Directly Induces an Involuntary Rapid Manual Following Response. *The Journal of Neuroscience*, 25(20), 4941–4951.
<https://doi.org/10.1523/JNEUROSCI.4143-04.2005>
- Schetzen, M. (1981). Nonlinear system modeling based on the Wiener theory. *Proceedings of the IEEE*, 69(12), 1557–1573.
<https://doi.org/10.1109/PROC.1981.12201>
- Schira, M. M., Fahle, M., Donner, T. H., Kraft, A., & Brandt, S. A. (2004). Differential Contribution of Early Visual Areas to the Perceptual Process of Contour Processing. *Journal of Neurophysiology*, 91(4), 1716–1721.
<https://doi.org/10.1152/jn.00380.2003>
- Sekuler, A. B. (1990). Motion segregation from speed differences: Evidence for non-linear processing. *Vision Research*, 30(5), 785–795.
[https://doi.org/10.1016/0042-6989\(90\)90104-S](https://doi.org/10.1016/0042-6989(90)90104-S)

- Sekuler, A. B., & Bennett, P. J. (2001). Generalized Common Fate: Grouping by Common Luminance Changes. *Psychological Science*, 12(6), 437–444.
<https://doi.org/10.1111/1467-9280.00382>
- SERGEANT, J., OHTA, S., & MACDONALD, B. (1992). FUNCTIONAL NEUROANATOMY OF FACE AND OBJECT PROCESSING: A POSITRON EMISSION TOMOGRAPHY STUDY. *Brain*, 115(1), 15–36.
<https://doi.org/10.1093/brain/115.1.15>
- Seymour, K., Karnath, H.-O., & Himmelbach, M. (2008). Perceptual grouping in the human brain: Common processing of different cues. *NeuroReport*, 19(18), 1769.
<https://doi.org/10.1097/WNR.0b013e328318ed82>
- Shinomori, K., & Werner, J. S. (2003). Senescence of the temporal impulse response to a luminous pulse. *Vision Research*, 43(6), 617–627.
[https://doi.org/10.1016/S0042-6989\(03\)00009-9](https://doi.org/10.1016/S0042-6989(03)00009-9)
- Shinomori, K., & Werner, J. S. (2008). The impulse response of S-cone pathways in detection of increments and decrements. *Visual Neuroscience*, 25(3), 341–347.
<https://doi.org/10.1017/S0952523808080218>
- Sikder, J., Islam, M. K., & Jahan, F. (2024). Object segmentation for image indexing in large database. *Journal of King Saud University - Computer and Information Sciences*, 101937. <https://doi.org/10.1016/j.jksuci.2024.101937>
- Song, Y., & Baker, C. L. (2007). Neuronal response to texture- and contrast-defined boundaries in early visual cortex. *Visual Neuroscience*, 24(1), 65–77.
<https://doi.org/10.1017/S0952523807070113>
- Straub, D., & Rothkopf, C. A. (2022). Putting perception into action with inverse optimal control for continuous psychophysics. *eLife*, 11, e76635.
<https://doi.org/10.7554/eLife.76635>

- Straub, D., Schultheis, M., Koepl, H., & Rothkopf, C. A. (2023). *Probabilistic inverse optimal control for non-linear partially observable systems disentangles perceptual uncertainty and behavioral costs* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2303.16698>
- Sun, Z., Chen, Y.-J., Yang, Y.-H., Li, Y., & Nishida, S. (2025). *Machine Learning Modeling for Multi-order Human Visual Motion Processing* (arXiv:2501.12810). arXiv. <https://doi.org/10.48550/arXiv.2501.12810>
- Sun, Z., Chen, Y.-J., Yang, Y.-H., & Nishida, S. (2023a). Modeling of Human Motion Perception Mechanism: A Simulation based on Deep Neural Network and Attention Transformer. *Journal of Vision*, 23(9), 4894. <https://doi.org/10.1167/jov.23.9.4894>
- Sun, Z., Chen, Y.-J., Yang, Y.-H., & Nishida, S. (2023b, November 2). *Modeling Human Visual Motion Processing with Trainable Motion Energy Sensing and a Self-attention Network*. Thirty-seventh Conference on Neural Information Processing Systems. <https://openreview.net/forum?id=tRKimbAk5D>
- Sun, Z., Wang, R., & Luo, Z. (2021). Polynomial approximation based spectral dual graph convolution for scene parsing and segmentation. *Neurocomputing*, 438, 133–144. <https://doi.org/10.1016/j.neucom.2021.01.002>
- Tanaka, K. (1996). Inferotemporal Cortex and Object Vision. *Annual Review of Neuroscience*, 19(Volume 19, 1996), 109–139. <https://doi.org/10.1146/annurev.ne.19.030196.000545>
- Tangemann, M., Kümmerer, M., & Bethge, M. (2024). *Object segmentation from common fate: Motion energy processing enables human-like zero-shot generalization to random dot stimuli* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2411.01505>

- Tolhurst, D. J. (1973). Separate channels for the analysis of the shape and the movement of a moving visual stimulus. *The Journal of Physiology*, 231(3), 385–402. <https://doi.org/10.1113/jphysiol.1973.sp010239>
- Tolias, A. S., Keliris, G. A., Smirnakis, S. M., & Logothetis, N. K. (2005). Neurons in macaque area V4 acquire directional tuning after adaptation to motion stimuli. *Nature Neuroscience*, 8(5), 591–593. <https://doi.org/10.1038/nn1446>
- Turk-Browne, N. B., Jungé, J. A., & Scholl, B. J. (2005). The Automaticity of Visual Statistical Learning. *Journal of Experimental Psychology: General*, 134(4), 552–564. <https://doi.org/10.1037/0096-3445.134.4.552>
- Uchikawa, K., & Yoshizawa, T. (1993). Temporal responses to chromatic and achromatic change inferred from temporal double-pulse integration. *JOSA A*, 10(8), 1697–1705. <https://doi.org/10.1364/JOSAA.10.001697>
- Umino, Y., Solessio, E., & Barlow, R. B. (2008). Speed, Spatial, and Temporal Tuning of Rod and Cone Vision in Mouse. *Journal of Neuroscience*, 28(1), 189–198. <https://doi.org/10.1523/JNEUROSCI.3551-07.2008>
- Usher, M., & Donnelly, N. (1998). Visual synchrony affects binding and segmentation in perception. *Nature*, 394(6689), 179–182. <https://doi.org/10.1038/28166>
- Uttal, W. R., Spillmann, L., Stürzel, F., & Sekuler, A. B. (2000). Motion and shape in common fate. *Vision Research*, 40(3), 301–310. [https://doi.org/10.1016/S0042-6989\(99\)00177-7](https://doi.org/10.1016/S0042-6989(99)00177-7)
- Vacher, J., Launay, C., Mamassian, P., & Coen-Cagli, R. (2023). Measuring uncertainty in human visual segmentation. *ArXiv*.
- Volberg, G., & Greenlee, M. (2014). Brain networks supporting perceptual grouping and contour selection. *Frontiers in Psychology*, 5. <https://www.frontiersin.org/articles/10.3389/fpsyg.2014.00264>

- von der Heydt, R., Peterhans, E., & Baumgartner, G. (1984). Illusory Contours and Cortical Neuron Responses. *Science*, 224(4654), 1260–1262. <https://doi.org/10.1126/science.6539501>
- von der Heydt, R., Zhou, H., & Friedman, H. S. (2000). Representation of stereoscopic edges in monkey visual cortex. *Vision Research*, 40(15), 1955–1967. [https://doi.org/10.1016/S0042-6989\(00\)00044-4](https://doi.org/10.1016/S0042-6989(00)00044-4)
- Wagemans, J. (Ed.). (2015). *Oxford handbook of perceptual organization*. Oxford University Press.
- Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. *Psychological Bulletin*, 138(6), 1172–1217. <https://doi.org/10.1037/a0029333>
- Wallis, G. (2005). A spatial explanation for synchrony biases in perceptual grouping: Consequences for the temporal-binding hypothesis. *Perception & Psychophysics*, 67(2), 345–353. <https://doi.org/10.3758/BF03206497>
- Watson, A. B. (1986). Temporal sensitivity. *Handbook of Perception and Human Performance*, 1(6), 1–43.
- Werkhoven, P., Snippe, H. P., & Alexander, T. (1992). Visual processing of optic acceleration. *Vision Research*, 32(12), 2313–2329. [https://doi.org/10.1016/0042-6989\(92\)90095-Z](https://doi.org/10.1016/0042-6989(92)90095-Z)
- Wertheimer, M. (1923). Laws of organization in perceptual forms. First published as Untersuchungen zur Lehre von der Gestalt II. *Psychologische Forschung*, 4, 301–350.

- Wertheimer, M. (1938). Laws of organization in perceptual forms. In W. D. Ellis (Ed.), *A source book of Gestalt psychology*. (pp. 71–88). Kegan Paul, Trench, Trubner & Company. <https://doi.org/10.1037/11496-005>
- Westheimer, G. (2004). Center-surround antagonism in spatial vision: Retinal or cortical locus? *Vision Research*, 44(21), 2457–2465. <https://doi.org/10.1016/j.visres.2004.05.014>
- Whitney, D., & Leib, A. Y. (2018). Ensemble Perception. *Annual Review of Psychology*, 69(Volume 69, 2018), 105–129. <https://doi.org/10.1146/annurev-psych-010416-044232>
- Williams, J. M., & Lit, A. (1983). Luminance-dependent visual latency for the hess effect, the pulfrich effect, and simple reaction time. *Vision Research*, 23(2), 171–179. [https://doi.org/10.1016/0042-6989\(83\)90140-2](https://doi.org/10.1016/0042-6989(83)90140-2)
- Wilson, H. R. (1993). Nonlinear processes in visual pattern discrimination. *Proceedings of the National Academy of Sciences*, 90(21), 9785–9790. <https://doi.org/10.1073/pnas.90.21.9785>
- Wilson, H. R., & Wilkinson, F. (1998). Detection of global structure in Glass patterns: Implications for form vision. *Vision Research*, 38(19), 2933–2947. [https://doi.org/10.1016/S0042-6989\(98\)00109-6](https://doi.org/10.1016/S0042-6989(98)00109-6)
- Wolfson, S. S., & Landy, M. S. (1999). Long range interactions between oriented texture elements. *Vision Research*, 39(5), 933–945. [https://doi.org/10.1016/S0042-6989\(98\)00144-8](https://doi.org/10.1016/S0042-6989(98)00144-8)
- Wong, Y. J., Aldcroft, A. J., Large, M.-E., Culham, J. C., & Vilis, T. (2009). The Role of Temporal Synchrony as a Binding Cue for Visual Persistence in Early Visual Areas: An fMRI Study. *Journal of Neurophysiology*, 102(6), 3461–3468. <https://doi.org/10.1152/jn.00243.2009>

- Wörgötter, F., Suder, K., & Fünke, K. (1999). The dynamic spatio-temporal behavior of visual responses in thalamus and cortex. *Restorative Neurology and Neuroscience*, 15(2–3), 137–152. <https://doi.org/10.3233/RNN-1999-00122>
- Wu, H., Zuo, Z., Yuan, Z., Zhou, T., Zhuo, Y., Zheng, N., & Chen, B. (2023). Neural representation of gestalt grouping and attention effect in human visual cortex. *Journal of Neuroscience Methods*, 399, 109980. <https://doi.org/10.1016/j.jneumeth.2023.109980>
- Xu, Y., & Chun, M. M. (2007). Visual grouping in human parietal cortex. *Proceedings of the National Academy of Sciences*, 104(47), 18766–18771. <https://doi.org/10.1073/pnas.0705618104>
- Yamane, Y., Carlson, E. T., Bowman, K. C., Wang, Z., & Connor, C. E. (2008). A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nature Neuroscience*, 11(11), 1352–1360. <https://doi.org/10.1038/nn.2202>
- Yang, X. L., & Wu, S. M. (1991). Feedforward lateral inhibition in retinal bipolar cells: Input-output relation of the horizontal cell-depolarizing bipolar cell synapse. *Proceedings of the National Academy of Sciences*, 88(8), 3310–3313. <https://doi.org/10.1073/pnas.88.8.3310>
- Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, 81(1), 141–145. <https://doi.org/10.1037/h0027474>
- Zacks, J. M., & Swallow, K. M. (2007). Event Segmentation. *Current Directions in Psychological Science*, 16(2), 80–84. <https://doi.org/10.1111/j.1467-8721.2007.00480.x>
- Zeile, A. J., Feigl, B., Kambhampati, P. K., Aher, A., McKeefry, D., Parry, N., Maguire, J., Murray, I., & Kremers, J. (2017). A Temporal White Noise Analysis for Extracting the Impulse Response Function of the Human Electroretinogram.

Translational Vision Science & Technology, 6(6), 1.

<https://doi.org/10.1167/tvst.6.6.1>

Zhou, H., Friedman, H. S., & Heydt, R. von der. (2000). Coding of Border Ownership in Monkey Visual Cortex. *Journal of Neuroscience*, 20(17), 6594–6611.

<https://doi.org/10.1523/JNEUROSCI.20-17-06594.2000>

Zhou, J., Benson, N. C., Kay, K. N., & Winawer, J. (2018). Compressive Temporal Summation in Human Visual Cortex. *The Journal of Neuroscience*, 38(3), 691–

709. <https://doi.org/10.1523/JNEUROSCI.1724-17.2017>