

修士論文

環境センサによる会話状況認識と
参加者による実時間タグ付けを用いた
ミーティングキャプチャシステム

指導教員 西田 豊明 教授

京都大学大学院情報学研究科
修士課程知能情報学専攻

中沢 拓磨

平成 21 年 2 月 6 日

環境センサによる会話状況認識と参加者による実時間タグ付けを用いたミーティングキャプチャシステム

中沢 拓磨

内容梗概

本研究では参加者 5, 6 人がプレゼンテーションとフリーディスカッションを交えて行うミーティングを対象として、その様子を映像と音声で記録するとともに、参加者の視線集中、写真撮影と電子メモの記入、プレゼンテーションスライドの操作などを検出してインデクスを生成することで、後から振り返るときのための利便性を確保したミーティングキャプチャシステムを構築する。このインデクスは映像の自動推薦を行う際に利用することを想定しており、過去をうまく想起させるような部分や、複数映像の中のふさわしい部分を選択するための指標となるようにする。収録したデータを自発的に振り返るときにもこのインデクスが手がかりとして提示され、見る人に対してどこから見ればよいのか、どのあたりが重要なシーンであるかが視覚的にわかるようにする。

視線集中の獲得は、参加者の顔方向および視線方向をミーティングスペース中央の全方位カメラによって観測し、東西南北のうち特定の方向に対して過半数の集中が起きたときを意味のある重要なシーンが起こったと推測する。この際、顔方向と視線方向は独立に判定される。顔方向集中と視線方向集中では顔方向集中のほうがより印象的な会話イベントが起こったことが示唆され、その特徴としては笑いやジャスチャを伴うことや、会話が質問応答の形式になっていることが挙げられる。

また参加者は捉えておきたい場面に対して、カメラのシャッターを押したりメモを書いたりする行為によって実時間タグと呼ぶインデクスを意識的に付与することが可能である。実時間タグのインデクスでは画像が生成されてから最後にメモ書きが起こるまでの活動時間に注目することにより、話題の継続性への手がかりとすることができる。

この両者は同時刻にはあまり発生せず、両者のインデクスは性質が異なる相補的なものである。そのため両者のインデクスは目的によって使い分ける必要がある。

Meeting Capture System Using Conversational Situation Recognition with Environmental Sensors and Real-time Tagging by Participants

Takuma NAKAZAWA

Abstract

In this paper, I construct a capture system for meetings, in which 4 to 5 participants make presentations and/or free discussions. It enables participants and none-participants to look over captured meetings. This system captures paging of presentation slides, pictures by PhotoChat and notes as indices in addition to videos and audios. Besides captured indexes, gathering of gazes are calculated and used as index. I suppose these indexes are good marks to recall meetings and to select a video from multi view point and they will be variable in auto-recommendation. They are also used as markers to show where to start to look and where to look for their importance when users want to look back.

A spherical vision camera set up at the center of meeting table captures meeting participant's faces in order to recognize that their gaze points gather at the same area. When the majority of them gather their gaze in any area of north, south, east, and west, the system guesses that a significant event occurs and generates the index. The precision ratio of the judgment of gaze gathering is about 65%. Some features are found in the conversations around the time that gaze gathering occurs.

Participants also can give real-time tags to the scene for which they want to leave the indices intentionally by pressing the shutter button and writing the memos. The length of time from the memo book daily life and happening of real time at the end can be used as a clue to the continuance of the topic.

The length of the time between the generation of the image of real-time tag and the last memo can be used as a clue to the continuance of the topic. Both are not so generated at the same time, and the indices of both is the complementary one another. Since the indices of both differ, it is necessary to use them properly by the purposes.

環境センサによる会話状況認識と参加者による実時間タグ付けを用いたミーティングキャプチャシステム

目次

第 1 章	はじめに	1
1.1	研究の背景	1
1.2	研究の目的	2
1.3	関連研究	3
1.4	本論文の構成	5
第 2 章	ミーティングを構成する要素	6
2.1	ミーティングキャプチャに求められる要素	6
2.2	ミーティングに現れる現象	7
2.3	キャプチャ対象となるミーティングの構成	8
2.3.1	カメラで捕らえるべき範囲と映像選択	9
2.3.2	実時間タグ	10
第 3 章	ミーティングキャプチャシステムの開発	12
3.1	システム構成と接続関係	12
3.2	カメラアレイと実装方法	14
3.2.1	カメラの型式	14
3.2.2	カメラの接続方式	14
3.2.3	カメラの同期方式	16
3.2.4	カメラの操作方式	16
3.2.5	映像のエンコード方式	16
3.3	全方位カメラと視線情報の処理	17
3.3.1	前処理	17
3.3.2	顔認識	17
3.3.3	顔追跡	20
3.3.4	3次元座標の復元	22
3.3.5	視線集中の検出	23
3.4	スライド操作記録ソフトウェアの実装	24
3.5	実時間タグ生成のための PhotoChat ログの処理	24

3.6	時系列インタフェース	24
3.6.1	ビジュアルによるインデックスの利用	25
3.7	文字列による検索	25
第4章	ミーティングへの適用事例	27
4.1	視線集中のインデックス	27
4.1.1	顔方向集中	27
4.1.2	視線方向集中	30
4.2	実時間タグによるインデックス	32
4.3	実時間タグと視線集中の比較	33
第5章	おわりに	35
5.1	結論	35
5.2	今後の課題	35
	謝辞	37
	参考文献	38

第1章 はじめに

1.1 研究の背景

日々行われる会話シーンで起きた出来事やその会話内容について，時間や場所を異にする第三者にそのまま伝えたいと思うことや，自分で後から見返したいと思うことはよくある．なぜならば会話中に起きた些細な行動や，細かな言い回しの中にも，重要な情報が含まれており，それらをうまく伝えることで，文脈を把握することが容易になるからである．このようなことを可能にするためには，会話をそのまま捉えることができるキャプチャシステムが必要である．

また会話が長期にわたる場合などは，手がかりなしに記録を観覧することは煩わしいものであり，再利用するためには映像全体を見返す必要があるとすれば，その魅力は損なわれることになる．例えば図1のような一覧表示を行うと，ほぼすべての画像が類似したものとなり，まったく見分けはつかない．よってユーザにとって必要な部分をかいつまんで観覧できるようにすることが望まれる．このために，発話内容を文字列として書き起こしておき，それらを手がかりにすることはしばしば行われる．しかし音声から議事録を書き起こすには人手を用いると大変なコストがかかり，コンピュータを用いた音声認識においても，高い精度を要求するためには音声波形データが理想的な品質である必要があるなど，条件の厳しいものとなっている．そこで言語情報の獲得コストを下げつつ，また，言語的な内容にとどまらず，視線やジェスチャのような非言語情報や，写真サムネイルを用いるなどの方法により，情報を容易に見つけることが可能なよう工夫が必要である．

さらに実際には重要であるにもかかわらず，その情報の存在に気がつかないために，零れ落ちていつてしまっている知識を伝達可能にするため，受動的に利用可能な記録の提示システムが必要である．例えば，ユーザに関係のありそうな過去の会話シーンのうち，自らが不参加だったものについて自動的に推薦を行ってくれたり，現在の会話の状況と関連するような過去の会話シーンを自動的に検索してくれたりするようなシステムが理想的である．

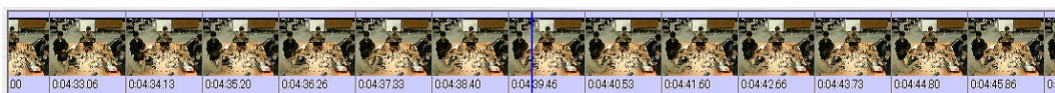


図 1: カメラ映像からの連続画像

1.2 研究の目的

本研究の目的は、参加者 4, 5 人がプレゼンテーションとフリーディスカッションを交えて行うミーティングを対象として会話のキャプチャを行い、それを観覧可能、再利用可能なものとするために、インデクスやキーワードなどのメタ情報が付与されたミーティングデータを半自動的に生成することである。

そのためにまず、同期された多視点映像を用意する。これは発話者の表情を的確に捉えた映像や、複数人の掛け合いによる会話など、ひとつの固定的な映像だけでは追いきれないものについても、データとして残していくためである。そのためにカメラを複数台用意することになるが、ユーザが映像を確認するための指標として、視線集中が起きている範囲を捕らえた映像がどれかを提示することが望まれる。

次にミーティング中の重要と思われる箇所にインデクスをつけるために、また視線集中が起きている領域を判定するために、参加者の視線方向を検出する。これらによるインデクスは人間の感性により近いと考えられ、関心の度合いなどと関連付けることにより、再利用性のある会話インデクスとなることを期待している。ただし収録の際はウェアラブルデバイスのような装着に時間がかかり、身体を拘束する可能性のあるものは使用しない。参加者の視線集中の検出にはテーブルトップの全方位カメラと画像処理方式を組み合わせる。これによりモーションセンサの装着などの手間なしにミーティングを始めることができる。画像処理の精度については完全ではないため、これを補完する方法についても検討する。参加者による手書きメモを生成するものとして、ネットワーク接続されたカメラ付小型タブレットコンピュータデバイスを使用し、その操作によって参加者の注目や関心の内容を表現する。これを参加者全員で共有することで、協調的にミーティングへの目印をつけていくことが可能である。またプレゼンテーションスライドの操作についても記録を行う。

最終的にはここで蓄積されたミーティングデータとメタ情報を元に、個々のユーザに対してミーティングの内容レベルに踏み込んだ自動推薦を行い、再利用を促す環境を構築することを目標とする。そのために、まず全方位カメラの画像から顔認識を行い、参加者氏名を抽出する。これによりミーティングごとの参加・不参加の状況を獲得できる。次にミーティングで使われているプレゼンテーションスライドの文字情報を抽出する。これは議事録の書き起こしを人

手で行うことに比べれば情報量は少なくなるが、要点が簡潔にまとまった情報であり、獲得コストも人手を解さないため低いものとなる。

想定するミーティングデータの再利用の流れを図2に示す。ここではミーティングの場から離れて過去のミーティングを見返すという用途のほかに、ミーティングの最中にも現在のミーティングと関連する過去の情報を引き出してくるという使い方を示している。例としては、ミーティングの最初の時間に、前回のミーティングのハイライトを表示することが挙げられる。他に、現在のプレゼンテーションスライドの言語情報からキーワードマッチングを行うということが考えられる。これらの場合、重要シーンや、次回までの課題などに関するシーンを表示できれば好ましい。

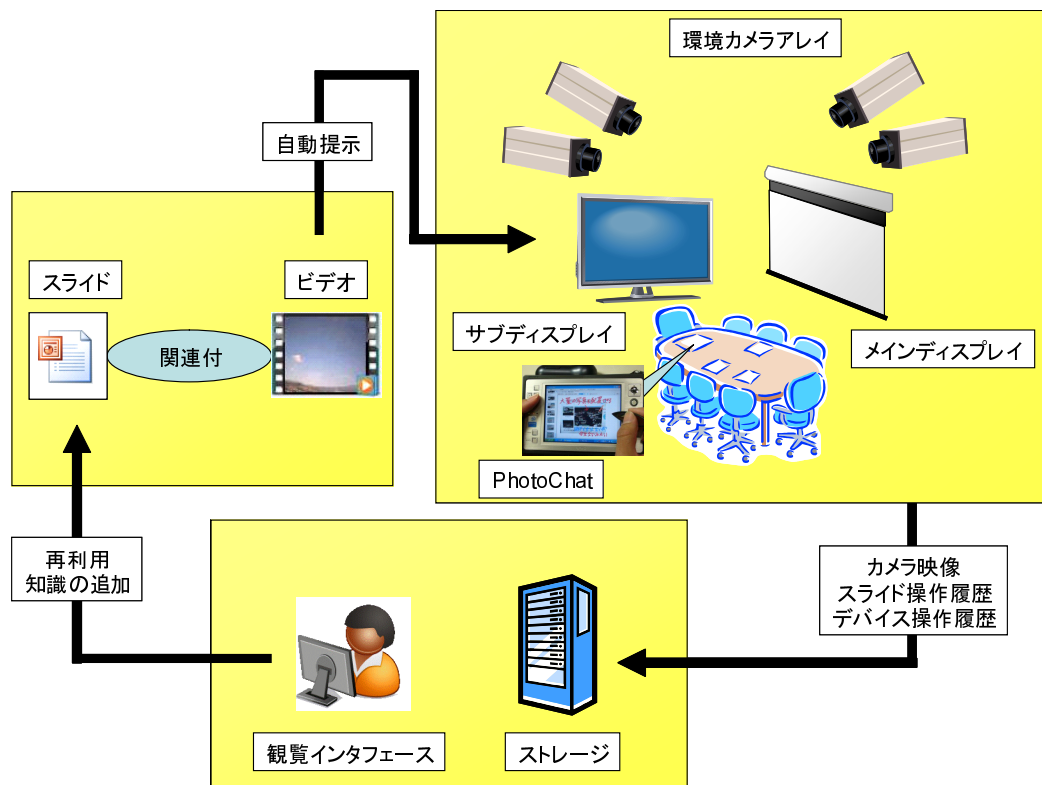


図 2: 再利用の流れ

1.3 関連研究

ミーティングを映像で撮り続けることにより、映像の特徴変化からシーンの切り替えを自動判別して、インデックスを付ける研究が従来からなされている [1, 2] 。

これらは会議の外面的特徴を利用したものであると言えるのに対し，本研究では個々の関心といった内面的特徴を同時に利用することが狙いである．

AMI[3] プロジェクトでは，ミーティング形式の会話について，会話の場をカメラやマイクロフォン等を用いて大量に記録・分析し，それらのデータをコーパスとしてブラウザとともに外部に提供している．このプロジェクトでは特に音声データに着目し，発話区間決定や音声認識のための基盤構築を目指している．

手作業で様々なモダリティのインデックスをつけることによって，ミーティングのコーパスを作成したものとして，VACE[4] が挙げられる．これに伴うインデクシング作業のコストは高く，日常的に行うことはできない．このためミーティングに参加しながらにして，参加者全員が協調してミーティングの注目点を選び出していくことで，日常的に使えるものを目指す．

ミーティングの再利用を目指した研究としてディスカッションメディア [5] が挙げられる．ディスカッションメディアは議論の構造を明確にすることで，高い検索性を有する議事録システムである．本研究の特徴は議論のセグメンテーションを発話など表だった行動だけでなく，内面的な要素をとまなうデバイスの操作からも行うことで，その内容をミーティングの場に還元し，参加者が受動的に知識の流れを得られることを目的としている．

レクチャー形式の講義の記録システムとして Classroom 2000[6] が挙げられる．講義映像，音声，スライド，板書などを記録し，学習者が利用可能なシステムを構築している．さらに講義アーカイブシステム [7] では講師と学習者の間の対話を記録可能であり，学習者のみならず講師も利用可能なシステムとなっている．

参加者の視線を自動抽出して，ミーティング映像にインデックスを付ける研究 [8, 9] がある．我々は同様に視線情報を利用するとともに，デバイスの操作を利用することで個々の注目を抽出することで，映像を再利用可能なものとする．

ミーティング内の会話を対象として，参加者の意思によって記録をつけていく方法として会話量子化 [10] が提案されている．この中でボタンデバイスによる記録が示されているが，本研究ではこれをカメラのシャッターボタンとして置き換え，さらに参加者の写真への書き込みを利用する．

1.4 本論文の構成

本論文は本章を含め5つの章から構成される。まず、第2章で本研究の問題設定について議論する。その上で、キャプチャ対象とするミーティングの性質を説明する。続いて第3章では、本システムの仕様と実装を述べ、システムを構成するコンポーネントを紹介する。さらに第4章では、本システムを適用したミーティングについて分析を行い、生成されたインデックスの使用可能性について述べる。最後に第5章で本研究の成果を総括し、今後の課題について述べる。

第2章 ミーティングを構成する要素

2.1 ミーティングキャプチャに求められる要素

ミーティングでは一般に，関係者が集まって相談をし，意思疎通や意思決定を行う．より細かな目的としては

- 報告
- 分析
- 課題解決
- 情報収集
- 意見収集
- 発想支援
- 交渉
- 交流

などが挙げられる．どのような目的のミーティングであれ，議論の過程や結果を記録しておくことは意義があるが，それぞれ重点的に記録すべき点は異なる．報告会においては進捗の状況や目標達成の度合いそのものが重要であるし，交流会においては誰とどれくらい話ができたのかが重要である．

ミーティングの記録の目的は，思い出すためと，他人に伝えるためである．このことから記録すべき点を列挙すると以下ようになる．

- いつ行われたか
- 参加者は誰か
- 本来参加すべきであったが不参加であったのは誰か
- 開催の案内はどのようになされたか
- ミーティングの種類は何か
- アジェンダは何か
- 参加者の役割は何か
- 誰が何をどれだけの時間だけ発言したか
- 発言に伴う態度はどのようなものか
- 議題はどう移り変わったか
- どのような資料を用いたか
- 資料にない情報は何か
- 何が新しく決まったか

これらの多くは議事録や映像記録によって取り扱えるものである．議事録の生成には人的リソースが必要であり，手間がかかるため，他の代替手段や，自動化によって解決されることが望まれる．

またそれらの記録があるだけでは必要な部分にたどり着くまでに多分に時間を要してしまう．そのため，同時に記録に対する索引が必要となる．議事録の場合は文字列による検索インデックスが作成できる．それだけではなく，ミーティングを通した移り変わりがどのようなになっているかを知るためには，大見出しや小見出しのように階層わけがなされていると把握しやすい．このような索引があれば議事録にも映像記録にも適用できる．

また多くの場合はプレゼンテーションスライドや手元資料が用意されるため，これらとミーティングの進行の対応関係を示すことは有用である．これらの資料はあらかじめまとめられているため，ミーティングの各部への索引として用いることができる．

これらから作られるインデックスは大きく二分できる．ひとつは重要シーンインデックスであり，もうひとつはシーン切り替わりインデックスである．シーン切り替わりインデックスは映画や音楽におけるチャプターに相当するものである．それらによって区切られた区間は一定の意味的まとまりを持つものとして定義される．それらの区間のなかに突発的に重要シーンが起こるものと考えられ，それが起こった時刻が重要シーンインデックスとなる．

2.2 ミーティングに現れる現象

ミーティング中の行動としては，発話のやりとりを基本としていくつかのものがある．一般的なミーティングは着席して行うものが多く，大きな移動が伴うことは少ないと考えてよい．行動については以下のものが挙げられる．

- 発話
- ジェスチャ
- 視線移動
- 資料へのメモ
- プレゼンテーションの開始

発話を捉えるには，ミーティングの際に音声を録音し，コンピュータによって自動音声認識を行うことで議事録を生成するということが考えられる．これを精度よく実施するには，ノイズの少ない環境下で発言者がはっきりと発話す

る必要があり，他者との会話のオーバーラップが好ましくないなど，制限事項が多い．

ジェスチャを捉えるには，モーションキャプチャによる方法が一般的であるが，身体を一部拘束するような形になり，使用する前に毎回準備が必要であるなど，日常的に使うには課題も多い．

視線を捉えるには頭部を追跡する方法，赤外線を眼球に照射する方法，画像認識による方法などがある．頭部を追跡や赤外線による方法は，やはり頭部にセンサなどを装着する必要がある．画像認識による方法であれば，参加者への負担はなくなる．

2.3 キャプチャ対象となるミーティングの構成

プレゼンテーション式ミーティングは特定の部屋においてスライドを映すスクリーンと，スクリーンの前の机に 4,5 人が着席するものを想定している．この際必要なものは環境側で用意し，ユーザが手元に必要なものは自分の利用するコンピュータとカメラ内蔵の小型デバイスである．

まずユーザはミーティングの開始時にシステムに対して開始の合図を送る．これは各々の手元のコンピュータで送信することができる．ミーティングが開始した後は，参加者は手元のコンピュータおよび小型デバイスの操作と会話を行う．プレゼンテーションを行う人は主に手元のコンピュータでスライドを操作しながら，内容について述べる．聞き手は小型デバイスを用いてミーティングの場面を写真として切り取るとともに，写真に対して手書きの注釈をつけたり，他者の撮影した写真を見たりする．この際，参加者はこれらの記録が後で利用されることを考慮して，重要とおもわれるシーンに対して重点的に撮影行為や注釈行為を行うように心がける．プレゼンテーションを行う人はミーティングの最中に交代可能であり，スライドによって違いを認識する．カメラアレイは静的に配置され，その撮影範囲は人物とその周辺を捉えるもの，全体を人物の背後から捉えるものである．全方位カメラは机の中央に設置され，各参加者の視線方向を捕らえる．複数人による視線方向の集中が起こると，その方向とともにイベントとして記録する．これにより注目を集めた対象とそれが起こったシーンを前後の文脈とともに再利用することが可能となる．

図3のように，プレゼンテーション式ミーティングが行われる場全体を俯瞰するものとして，複数のカメラを用いて多方向から同期的に収録する．話し手は

スライドを用いてプレゼンテーションを行うが、この際スライドに現れる言語的内容とスライドの操作履歴をシステムが捕捉する。プレゼンテーションの聞き手はカメラ型キャプチャデバイスである PhotoChat[11, 12] を用いてミーティングに参加する。

これらの情報を関連付けてストレージに保存し、それぞれの情報を統合して観覧可能な状態にする。ユーザはこれらの情報から重要な点を得ることができる。

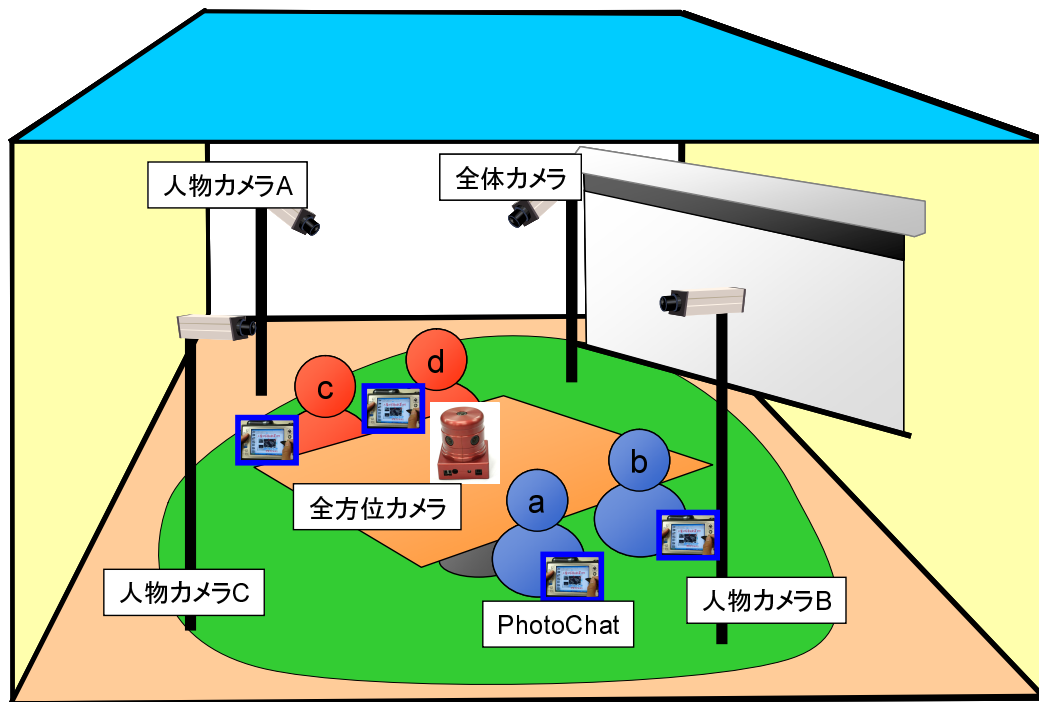


図 3: ミーティング空間の概要

2.3.1 カメラで捕らえるべき範囲と映像選択

まずプレゼンテーション式ミーティングを撮影の対象としていることから、スクリーン方向のカメラとスクリーンから全体を見渡すカメラを用意する。

次に人物に向かう方向の映像が必要である。この際撮影範囲は同じ方向に座る人物が全員含まれるようにする。本研究では長方形の机を利用するため、人物カメラは4台用意する。さらに各個人をより近くで捕らえるカメラを設置し、表情や視線、口の動きなどを見ることができる映像を捉える。

このようなカメラの設置状況の下で、映像選択をすることを考える。参加者はそれぞれ手元に PhotoChat デバイスを持ち、これを用いてミーティングのシー

ンを切り取ってゆくのであるが，積極的な発話姿勢にある人は，PhotoChat を同時に使うことは難しい．すなわち積極的な発話姿勢では，目配せをし，ジェスチャを伴っていると考えられるので，PhotoChat を見ることも持つこともできない状況である．このため，PhotoChat の利用頻度の低いところに発話の中心があると考えられる．スライドの状況から，発表者がどの人物であるかは判定することができるが，発話者は判定できない．

視線集中が行われる対象は，発話者，発話を直接受ける対象者，発話内容に言及された対象物と考えられる．カメラが捉える領域によって，部屋の空間を大まかに分割し，その領域に対する視線集中を計測する．視線集中が行われている領域を写すカメラを選択することで，視線の対象物を映す映像となる可能性が高い．

2.3.2 実時間タグ

PhotoChat とは，カメラを内蔵した小型コンピュータ上で動作するソフトウェアで，デジタルカメラのように写真を撮影できるとともに，手書きペンを用いて写真に書き込みをすることができる．それらの写真や書き込みはネットワークによって他の参加者と共有されるので，その上で自然にチャットを行うことが可能である．これによりミーティングの内容に踏み込んだタグ付けがリアルタイムにかつ自然に行われることを期待している．

PhotoChat はカメラ付小型タブレット PC デバイスであり，このデバイス上で写真と手書き文字によって他のユーザと会話が可能である．以前より議事録作成ツールの可能性が指摘されているが，本格的な利用を目指すには客観的なカメラ映像との同時運用が望ましい．

図 4 のように，PhotoChat には議事録としての運用可能性がある．その例としては発言中の話者の写真を撮り，その写真に会話ふきだしを付与しそこに会話内容を記述するものがある．また話し手が一方的にプレゼンテーションを行っている場であっても，聞き手側はそれをさえぎることなくスライドを写真に取り，その上でメモや質問，内容へのつつこみなどの会話を行うことができる．このような聞き手側の行動はカメラやマイクで捉えきることができず，聞き手の内面の表現が PhotoChat 上で起こっていると考えられる．この聞き手の内面の変化は，カメラなどに映る外界の状況からもなんらかの影響を受けていると考えられる．そのため PhotoChat の操作履歴からカメラ映像に対して目印を付与したうえで，その利用方法を検討する．



図 4: PhotoChat の議事録利用

PhotoChat の特徴としては、ユーザが撮影対象に気がつきその重要性を計った上でカメラを構えて撮影するため、実際の注目時刻に対して撮影時刻が遅れる可能性がある。さらにその上の手書き文字が実際の発話内容を忠実に再現するものだとするれば、手書き文字に対応する発話はその時刻以前に行われたことになる。

別の可能性として、PhotoChat のユーザは随時その撮影・書き込み内容を確認するのであるが、特定の写真に対して全体の注目や書き込みが集中する時間については、なんらかの重要な部分が発話についても起こっていると推測できる。たとえば、スライドの特定の部分について深く言及が起こったり、時間的にさかのぼる内容に言及されたとき、ユーザはそのことを確認しようとしたり、思い出そうとして、それらが表現された写真、あるいは書き込みに注目しようとすると考えられる。そのためユーザの動きからミーティング全体のダイナミクスを取得して、映像の時間ごとの利用価値を判別することに役立てることができる。

第3章 ミーティングキャプチャシステムの開発

このような条件を満たすミーティングを記録・観覧するためのミーティングキャプチャシステムを開発した。システムが捉えるミーティングの要素は以下である。

映像および音声の収録 音声と同期した映像によって発話と発話周辺行為を視認することができる。これらは、ミーティングを振り返る際に再生される。
イベントおよび重要シーンの獲得 ミーティングの最中の重要シーンを捉える。
参加者の意識的に指定した重要シーンと、参加者の視線集中から結果的に重要となるシーンを推定する。

文字注釈 イベント内容を言語的に表した注釈である。

画像注釈 スライド、写真、映像のサムネイルなどからなる注釈である。

3.1 システム構成と接続関係

システム構成を図5に示す。ミーティングを捉える4種類のデバイスに入力されたデータは、いくつかの処理を経た後、時刻に紐付けられたイベントとして管理される。

デバイスの接続関係を図6に示す。ユーザが本システムを使用する際は、手元にコンピュータを用意し、ネットワーク経由でシステムにアクセスする必要がある。その際、インターネット経由・イントラネット経由、また有線経由・無線経由を選択できる。インターネット経由のアクセスはプライバシー保護の観点から一部制限されているため、イントラネット経由での使用が望ましい。またコンピュータの位置が自由になることから、無線経由での使用が望ましい。無線はWi-Fiを用いる。システムの心臓部はクラスタコンピュータシステムInTrigger[13]である(図7)。これにより大量のデータ処理が可能となる。カメラアレイ、PowerPoint、PhotoChatから出力される、映像・音声・画像データはストレージに保存される。それ以外の時刻データ・文字データなどのメタデータはデータベースへと格納される。

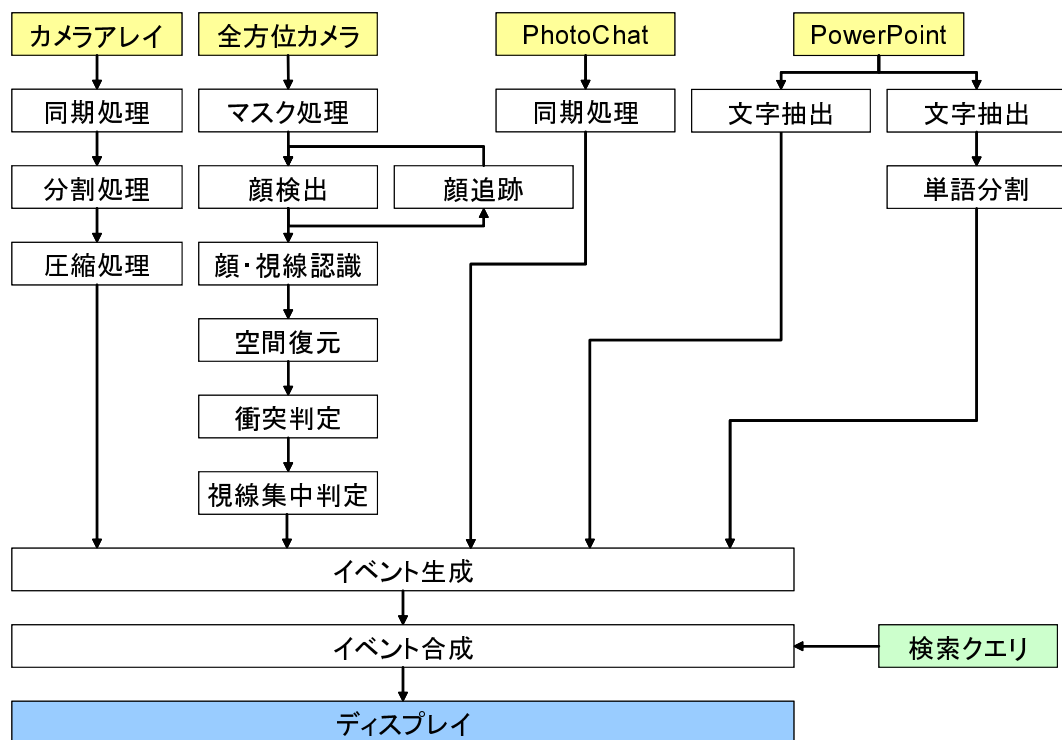


図 5: システム構成図

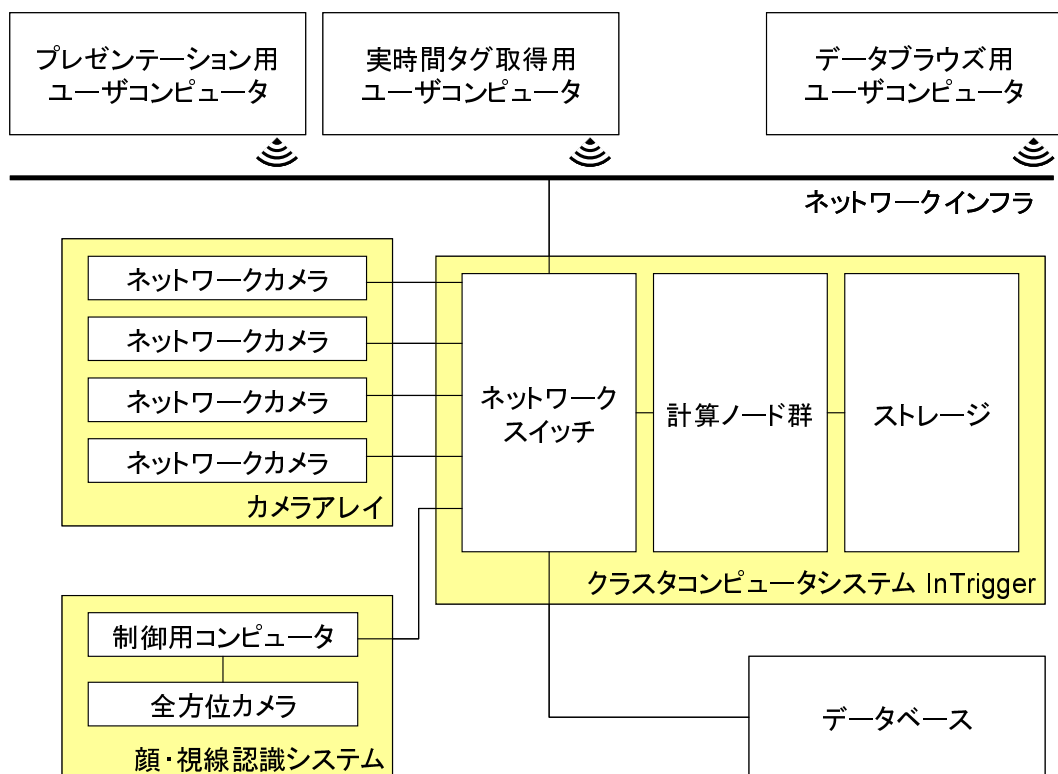


図 6: 接続関係図

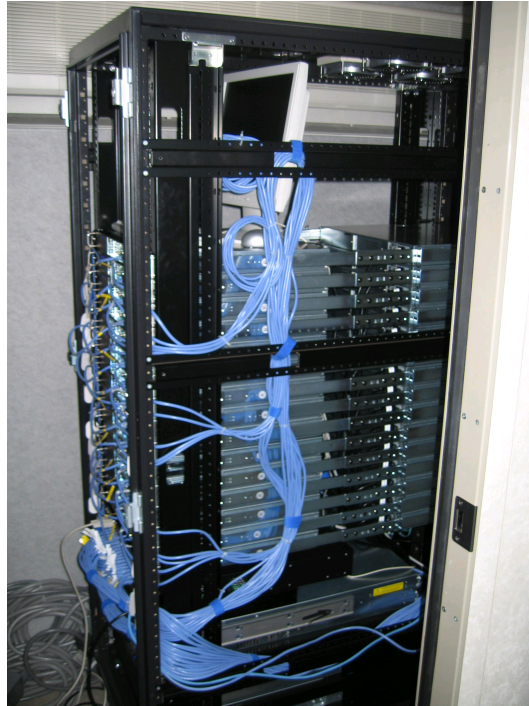


図 7: クラスタコンピュータ InTrigger

3.2 カメラアレイと実装方法

3.2.1 カメラの型式

このカメラアレイシステムは、コミュニケーションセンサルームである IMADE Room[14] に設置されている。これまでに IMADE Room を用いた多目的な実験用途にカメラアレイシステムは使用されている。現在は AXIS 210A(図 8) を 8 台、AXIS 214 PTZ を 1 台の計 9 台で構成されている。解像度は 640×480 であり、最大フレームレートは 30fps、ビットレートは 10Mbps 程度である。図 9 のような映像が得られる。

3.2.2 カメラの接続方式

カメラは Ethernet によるネットワーク機能を持つものを選択し、制御用コンピュータとの接続はこのネットワーク機能を介して TCP/IP によって行った。カメラは室内のいたるところに設置可能である必要があるが、カメラの接続方式が IEEE1394 や USB であると、ケーブル長が数メートルまでに制限される。そのためカメラの設置箇所のすぐ隣に制御用のコンピュータを配置する必要があるが、この場合それぞれのコンピュータはマスターの制御コンピュータに Ethernet などですらに接続する必要が出てくる。またカメラの制御に十分な性能を持つ



図 8: AXIS 210A

程度のコンピュータを，自由に配置されるカメラのすぐ隣に設置することは困難である．長いケーブルを用いることができることにより，制御用コンピュータは離れた位置に一括して設置されるため，カメラの設置の自由度を阻害しない．

制御用コンピュータはクラスタ構成のコンピュータを用いている．撮影の際は複数カメラを同時に作動させるため，複数の動画ストリームが同時に取り扱われることになるが，安定性を確保するため，コンピュータのノード1つに対してカメラ1つを割り当て，互いのストリームが干渉しないように構成している．

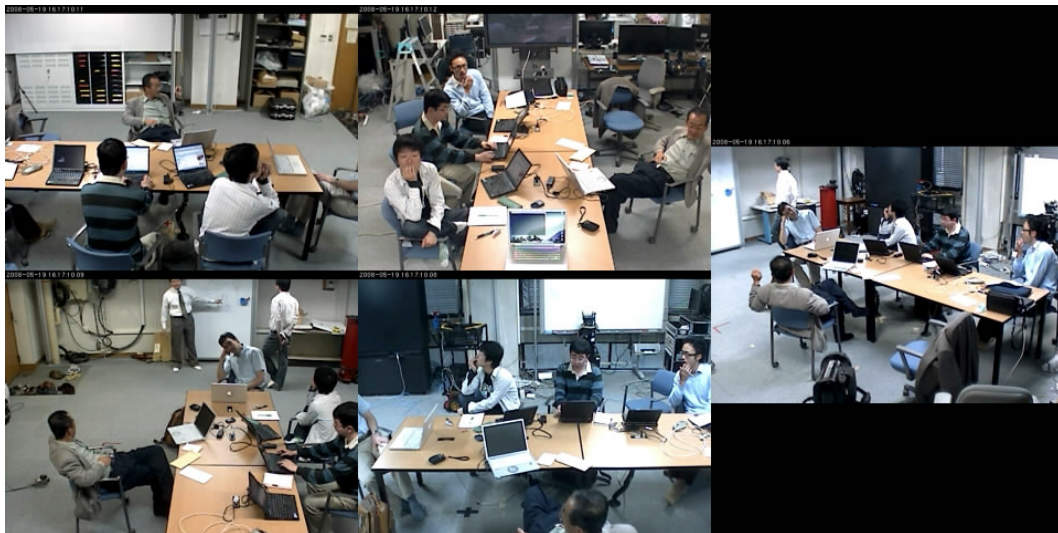


図 9: カメラアレイでミーティングを捉えた例

またクラスタノードをソフトウェア的に組み替えることができるため，クラスタノードに故障が発生した場合にも，別のクラスタノードを利用して記録が

できるようなシステムになっている．実際故障が発生し，複数のノードが使用不能になった際にも，接続定義ファイルを更新するだけですべてのカメラが利用可能となった．

3.2.3 カメラの同期方式

各カメラから出力される映像のフレームレートは負荷などによって変動する．これを補正するためには1フレームごとにすべてのフレームの調整を行う必要がある．それはフレームの変動による映像の時間とのずれは撮影時間がたつにつれて蓄積するため，部分だけを修正することはできないからである．本実装におけるシステムでは，ネットワークカメラ，制御コンピュータの内部クロックを，それぞれ制御マスターに対してNTPで同期させ，1フレーム単位に時刻情報を持つことで同期可能とした．各フレームの持つ時刻情報は制御コンピュータに送られた時刻ではなく，フレームが生成された時刻となる．映像の撮影開始シグナルと撮影終了シグナルの同期発行も実装されており，ファイル単位での比較も容易に行える．

3.2.4 カメラの操作方式

簡単に実装するには，カメラを起動する際にコンソールからコマンドを実行するような方式をとればよいが，カメラシステムのユーザの利便性を考慮し，Webインタフェースのボタンにより撮影開始と撮影終了を可能にしている．これはRubyを用いたCGIにより実装している．これにより，LAN内の任意のPCから操作が可能となり，操作者の位置の自由度も改善されている．

3.2.5 映像のエンコード方式

カメラから得られる映像ストリームはMotion JPEG形式であるが，Web上で取り扱うには適していない．そのため圧縮処理と形式変換を行う．蓄積された元映像に対して，定期的にエンコードイベントが発生し，クラスタによる並列計算によって高速にH.264符号形式とする．圧縮にはFFmpeg¹⁾を一部改良したもの²⁾を用いる．時刻情報は，tc2mp4³⁾とMP4Box⁴⁾を用いてタイムコードを生成し，映像コンテナであるMP4形式に埋め込むことで，可変フレームレート映像として整形する．これらの圧縮変換命令をクラスタの各ノードに発行する

¹⁾ <http://ffmpeg.org/>

²⁾ LGPL に従い再配布可能

³⁾ <http://d.hatena.ne.jp/zmi/>

⁴⁾ <http://gpac.sourceforge.net/>

ために、TORQUE¹⁾を用いている．また取り扱いの簡単さのために長時間の映像は分割してファイル化される．

3.3 全方位カメラと視線情報の処理

全方位カメラ Ladybug2 と顔認識ソフトウェア OMRON OKAO Vision によって視線方向を取得する．まず図 10 のように机上に全方位カメラを設置する．全方位カメラの高さは、机上のものが顔を遮らないように底上げを行い、また参加者間の視線のやり取りを阻害しないよう、顔よりやや低い位置に設置されている．

参加者の顔映像から画像認識を行うことで、おおまかな視線方向を取得する．ハードウェアの処理負担のバランスを考慮し、フレームレート 1.1fps で映像を取り出し、カメラのゆがみ補正をかけて視線認識を行い、席の配置のモデルから視線対象を推測する．特に複数人の視線の対象が一致する状況を取り出すことを目的とする．

3.3.1 前処理

全方位カメラから得られる映像は、図 11 のようなものになる．この映像から視線情報を取り出し、3 次元空間内の情報として構成する方法について述べる．

まず、顔認識を行う前に前処理として画像にマスクを適用する．マスクする範囲は画像上方（天井）および画像下方（机）である（図 12）．

これにより、顔の探索範囲を制限し、探索精度を向上することができる．実際、マスクを行わない場合では、机の色が肌色に近いので、色情報を用いた探索を行うときに問題が起こる場合があった．

3.3.2 顔認識

次に顔認識を行う．顔認識は複数の方法を混合して用いる．主たる顔認識手法として、OKAO Vision[15] を用いる．これにより、映像の 1 フレームごとに顔器官を含む矩形領域、顔方向、視線方向および顔氏名を取得する．OKAO Vision による認識が失敗したときは、CAMSHIFT[16] により顔座標を、カルマンフィルタを用いた時系列推定により顔領域を取得する．図 13 にこれらを示す．図の外側から順に、CAMSHIFT による顔追跡領域（赤）、カルマンフィルタによる推定顔領域（紫）、OKAO Vision による矩形領域（緑と赤）、顔方向（赤）、視

¹⁾ <http://www.clusterresources.com/>

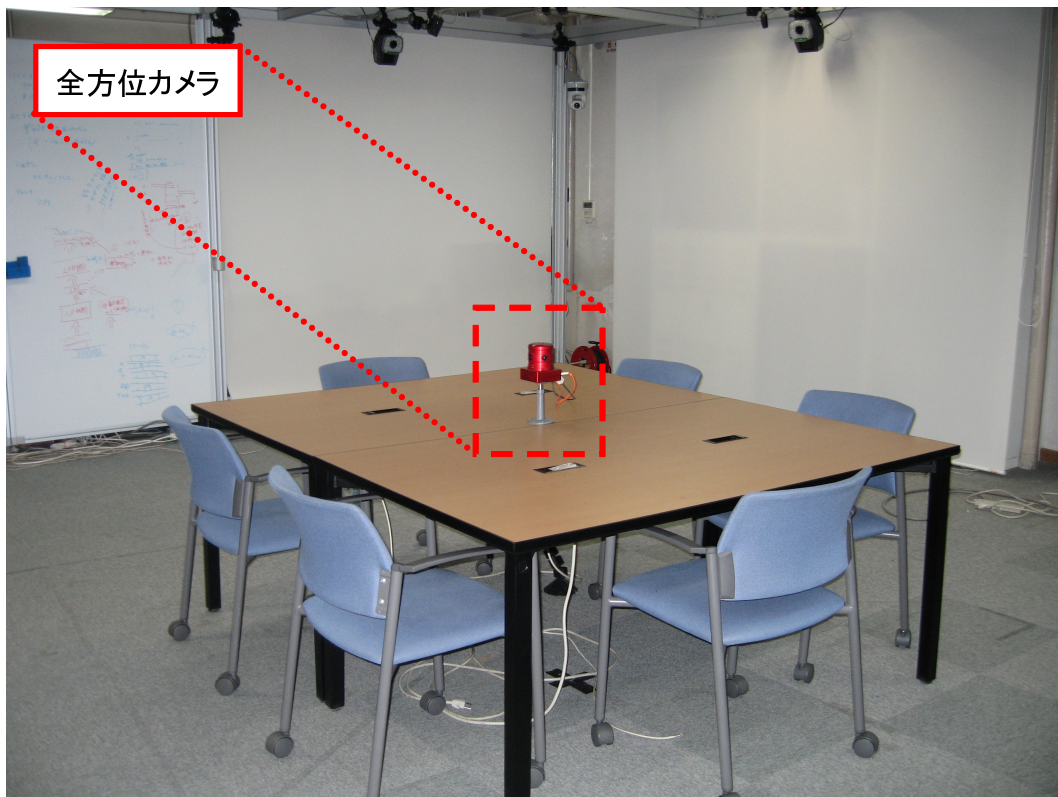
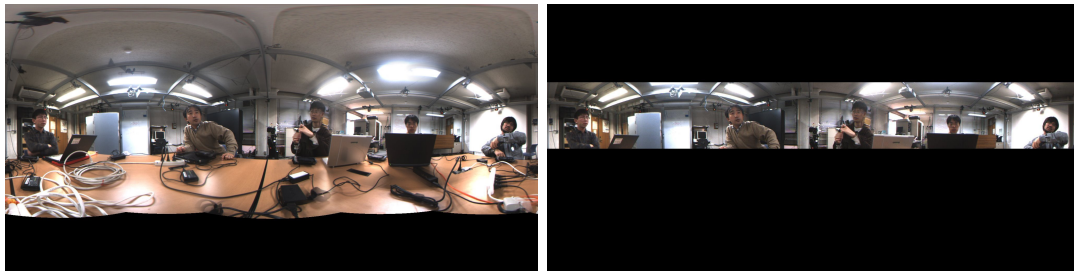


図 10: 全方位カメラの設置



図 11: 全方位カメラから得られる映像例



(a) マスク前

(b) マスク後

図 12: 映像のマスク

線方向（青）である．

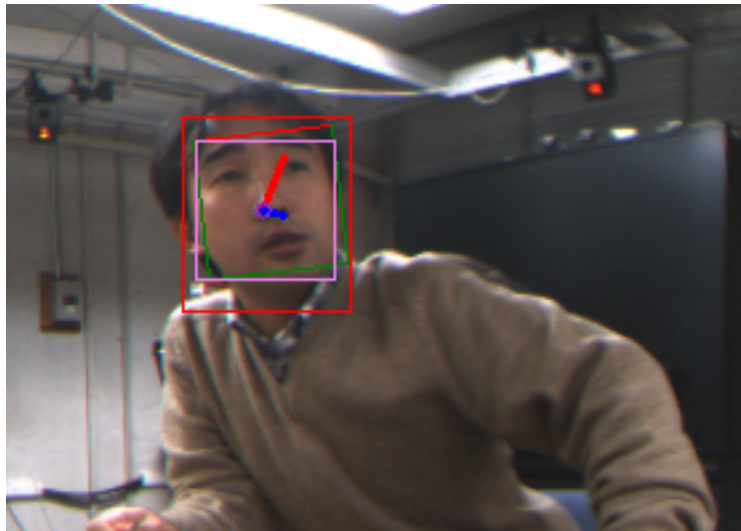


図 13: 顔認識

OKAO Vision の認識失敗が起こる要因として以下のものが挙げられる．

- 顔とカメラの間を手・PhotoChat 装置・ノートパソコンによって遮られる
- 顔の高速移動
- 顔が全方位カメラのレンズとレンズの境界にかかる
- うつむいて顔を下に向ける
- 撮影対象が起立する

これらが起こった場合，上記補間によって顔位置を取得する．

3.3.3 顔追跡

同一の顔に対して一意の ID を振る手法として，上記顔位置に関して 1 フレームごとに最近傍のものを追跡する手法と，OKAO Vision の顔識別を利用した顔氏名によるものを併用する．これは両者の誤検出の特性が異なることによる．

本研究で対象とするミーティングでは，参加者は椅子に着席し半固定された状態であるため，初期位置からの大きな移動については起こりにくいと考えて良い．そのため最近傍法を用いればほとんどの場合問題なく追跡することができる．この手法に問題が発生する場合として，顔が長時間障害物に隠れた場合が挙げられる．このとき OKAO Vision による顔検出はされず，CAMSHIFT とカルマンフィルタによる推定値のみが計算されるようになっている．そのため長時間，たとえば 240 秒程度認識不能な時間が続いたり，同時に隣り合う二人が認識不能になったりすると，他者の顔領域まで推定値が移動してしまい，結果として顔の ID が混ざり合ってしまう．この顔追跡の失敗を補正するため，顔氏名と顔領域の横方向への移動量を用いる．

まず顔氏名の誤検出を取り除くために， $N = 20$ フレームだけ過去の情報を取り出し，最頻の顔氏名を現在の顔氏名とする．これは顔氏名の誤検出が表 1 のようにパルス的に起こるからである．最頻値をとることで平滑化効果を生み，ノイズを除去できる．また認識不能が最頻であった場合には認識不能という顔氏名とする．

次に顔領域の横方向への移動量を計算する．直前のフレームが認識可能であった場合はそのフレームからの移動量を，認識不能であった場合は直近の認識可能であったフレームからの移動量を計算する．これが画像上で閾値 $th = 200$ ピクセル以上，すなわち実空間上で机中心からの角度が約 35 度以上ずれた場合， N フレーム後に顔 ID をリセットし，新たに顔氏名を取得して顔 ID とする．これはそれぞれの顔は移動しにくいという特性を利用し，大きく移動しすぎた場合は誤認識であると判定することに相当する．またリセットまでに N フレームの時間をおくのは，上記平滑化効果を得るためである．

この方法を用いると，稀に同一顔氏名が 2 つ以上同時に存在してしまうことがある．このとき，正しい顔氏名は高々ひとつであるから，その時点で認識成功している顔に対してのみ顔氏名を付与する．複数認識成功している場合とすべて認識失敗している場合は，判定順が先頭のをひとつ選ぶ．

表 1: 顔認識のパルスのエラーの例

フレーム番号	時間 [秒]	顔 ID	顔氏名	X 座標	Y 座標
769	697.26304	0	Alice	763	429
770	698.16975	0	Alice	763	429
771	699.07646	0	Alice	763	425
772	699.98318	0	Alice	761	427
773	700.88989	0	Alice	760	426
774	701.79661	0	Alice	781	426
775	702.70332	0	Bob error	788	426
776	703.61003	0	Eve error	791	423
777	704.51675	0	Alice	756	432
778	705.42346	0	Alice	757	433
779	706.33018	0	Alice	756	433
780	707.23689	0	Alice	759	433
781	708.1436	0	Alice	764	431
782	709.05032	0	Alice	761	434
783	709.95703	0	Alice	761	432
784	710.86375	0	Alice	762	433
785	711.77046	0	Alice	762	433
786	712.67717	0	Alice	763	433
787	713.58389	0	Alice	765	431
788	714.4906	0	Alice	765	429

3.3.4 3次元座標の復元

得られた情報から実際の3次元構造を復元することを考える．ミーティングは図14のような机を取り囲んで行われる．ここで参加者は机から250mm離れた，図14の点線部分に頭部中心を置くと仮定する．これは脇を閉めて椅子に座り，手を机の端に置いたとき，机の端から胴体中心までの距離がおよそ前腕長となることによる．この制約によって，2次元情報から3次元情報を復元することが可能になる．

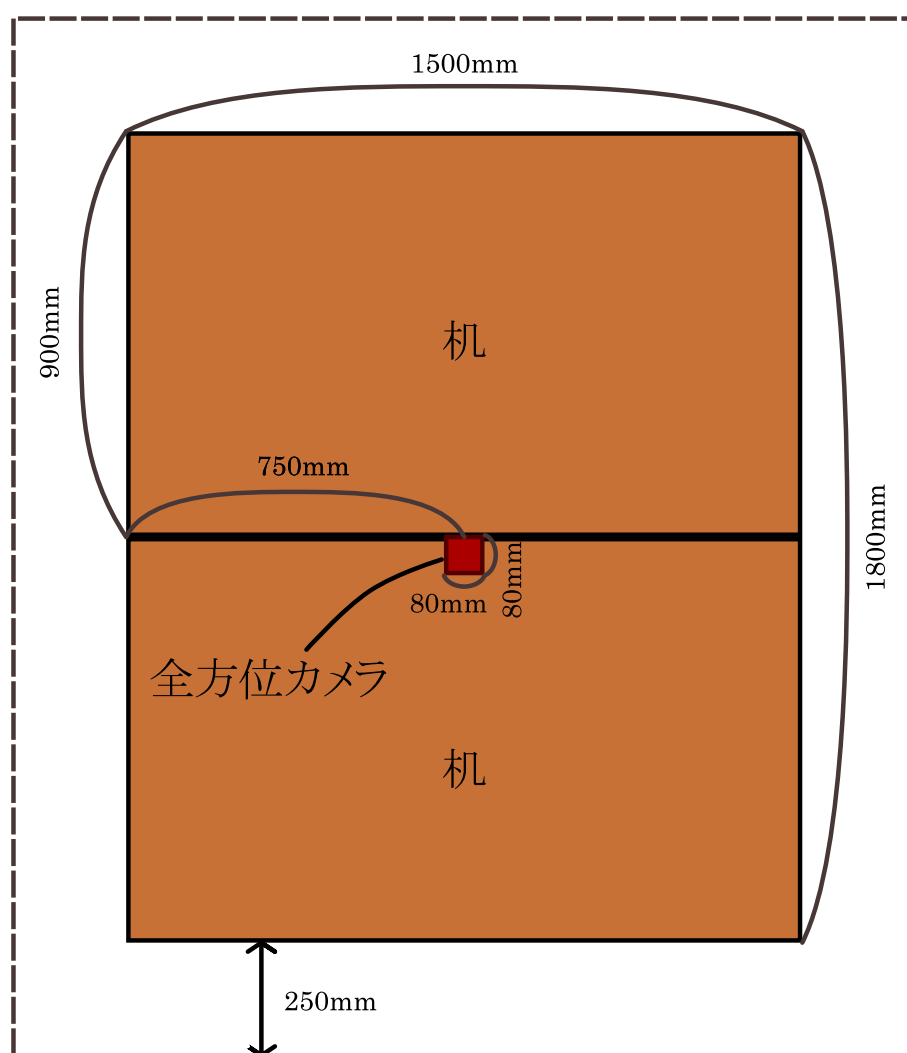


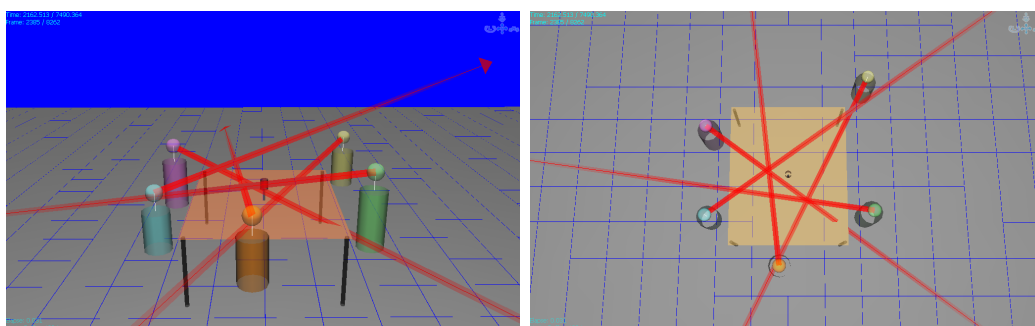
図14: 机，全方位カメラ，頭部中心位置

この方針で3次元座標を復元すると図15のようになる．以下に具体的な変換手法を述べる．まず図11の原画像は球面を2048×1024ピクセルの平面に放射

状マッピングしたものである．よって画像の x 座標はカメラ中心からの方位角 θ , y 座標は仰角 φ に線形変換できる．

$$\theta = \frac{x_0}{1024}\pi, \varphi = \frac{(512 - y_0)}{1024}\pi,$$

これにより平面内の顔位置を 3 次元の方向成分に変換し，この方向成分と図 14 の点線部分（面）との交点に頭部中心を生成する．さらに OKAO Vision による顔方向と視線方向の認識結果が左右角と上下角で得られるため，左右角 0 度，上下角 0 度となる方向を頭部中心からカメラ中心へのベクトルと設定し，同空間内の直線として生成する．



(a) 俯瞰図

(b) 上面図

図 15: 3 次元座標の復元例

3.3.5 視線集中の検出

OKAO Vision による視線方向の認識結果はそれほど精度が高くなく，また参加者の座標を机から 250mm と規定したことによる実際との誤差が生じるため，視線が特定の人物を貫いているかどうかという精度で判定することは難しい．そのため視線方向の分類を大雑把に行い，東西南北と机の 5 種類とする．

参加者のデフォルトの視線方向は，椅子に座った正面方向である．そのため，参加者人数 5 人の配置では何もしなくとも同じ方角に 2 人の視線集中が発生する．よって同じ方角に 3 人以上の視線集中が起こった時刻を検出する．また顔方向検出についても同様に行う．

3.4 スライド操作記録ソフトウェアの実装

ミーティングの参加者がプレゼンテーションスライドを用いるとき、この情報を再利用可能なようにサーバに収集するためのソフトウェアを開発した。このソフトウェアは現在最も一般的なプレゼンテーションソフトウェアである Microsoft PowerPoint のアドオンとして機能し、プレゼンテーションに使用するコンピュータに事前にインストールしておけば、どのようなコンピュータでも有効となる。その動作は、まずスライド内に配置された文字情報を取り出すとともに、スライド自体を画像情報として保存し、データベースと繋がるサーバへバックグラウンドで送信を行う。さらにスライドの切り替え操作が行われるたびにそのイベント情報をサーバへ通知する。このようにしておくことで、あとからこのミーティングのデータを再生する際にスライドの切り替え情報とスライドの画像を用いることができる。またサーバに送られた文字情報は、サーバ内で形態素解析ソフトウェア MeCab[17] を用いて単語単位に分割され、補助動詞を除いた動詞と数詞を除いた名詞のみがデータベースに送られる。スライド上の文字情報は情報が集約されたものと考えられるため、ミーティングの内容を検索する手がかりとなり得る。

3.5 実時間タグ生成のための PhotoChat ログの処理

参加者は PhotoChat を用いてミーティングの興味深いシーンを切り出すことができる。書き込みと書き込みの間が 10 秒以内の場合は同一イベントとみなす。

3.6 時系列インタフェース

得られた映像音声情報およびイベント情報は Web ページ上に表示され、一般的な Web ブラウザによって観覧できる。一般的な環境で観覧できることは、ユーザがミーティングを見返すにあたっての敷居を下げることに寄与する。実装は InTrigger 上に Ruby on Rails を用いて行った。インタフェースの例を図 16 に示す。左から右へと時間軸が流れ、イベントが起こった時間帯を示すとともに、イベントの内容を簡単なタイトルと、サムネイル画像によって表現している。操作環境 API として SIMILE Timeline[18] を用いており、動的な画面操作が可能である。画面操作の例としては時間のスクロール、時間軸の拡大縮小、イベントの観覧などがある。

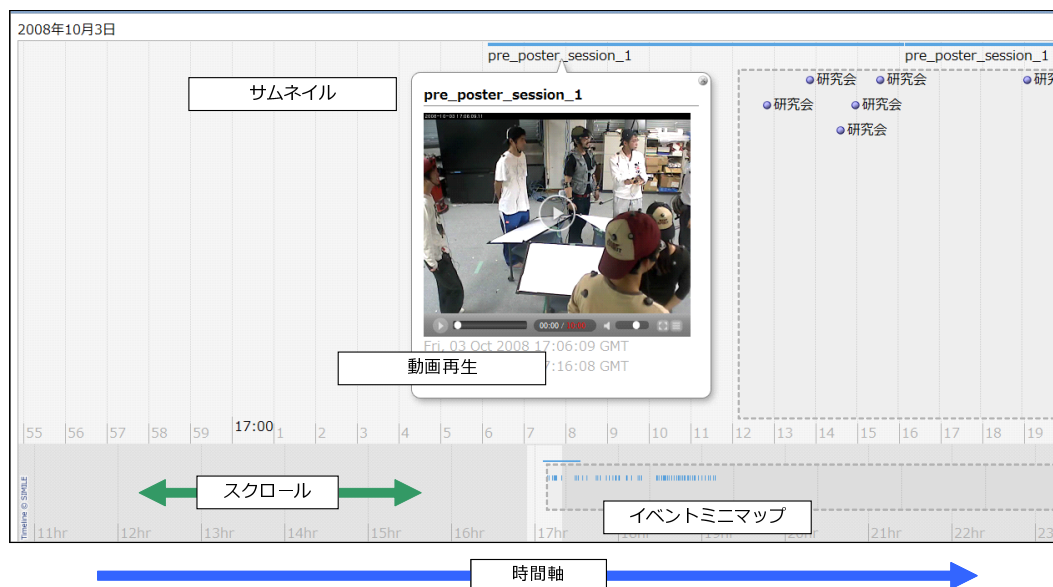


図 16: ブラウズインタフェースの例

3.6.1 ビジュアルによるインデックスの利用

サムネイル画像となる写真やスライドをどのように表示するかという問題がある。スライドイベントや撮影イベントは多くても1分に1回程度の頻度であるため、観覧インタフェースが時間軸の拡大縮小を可能なことを考えると、イベント自体をすべて表示することに問題はない。しかしサムネイルの大きさを考慮する必要がある。画面を埋め尽くしきらない範囲で、優先度をつける必要がある。一般的な方法としてはスライドを優先するとともに、書き込み頻度が高かった写真を優先する方法が考えられる。別の方法として、特定のユーザの撮影や書き込みを優先することも考えられる。これは特に自分自身のミーティング参加時の記憶を想起させやすくするといったことや、先生などの特に重要な意見を発するユーザの視点を追うといったことで使用できる。

3.7 文字列による検索

データベースに蓄積されたスライド上の文字情報を、単語単位で全文検索することが可能である。これはMySQLの全文検索機能を用いている。このとき、単語数の少ないキーワード、例えば1単語や2単語を用いると、そのキーワードを含むあらゆるスライドが検索結果として現れるため相応しくない。一方で、検索キーワードをスライド1枚分の単語すべてとするならば、それらの単語の

一部をより多く含み，そうでない単語がより少なく含まれるものから順に検索結果として現れる．これにより，スライドによってスライドを検索するということが可能となり，ひいては現在プレゼンテーション中のスライドと関連のあるスライドが使われたミーティングを検索することができる．

第4章 ミーティングへの適用事例

図 17 のように，参加者数 5 人が机の各辺に沿うような状況で 2 時間 4 分 50 秒のデータを収録した．ただしこの途中 46 分 7 秒間は全員が机から離れ別の場所に移動したため，有効な収録時間は 1 時間 18 分 43 秒である．

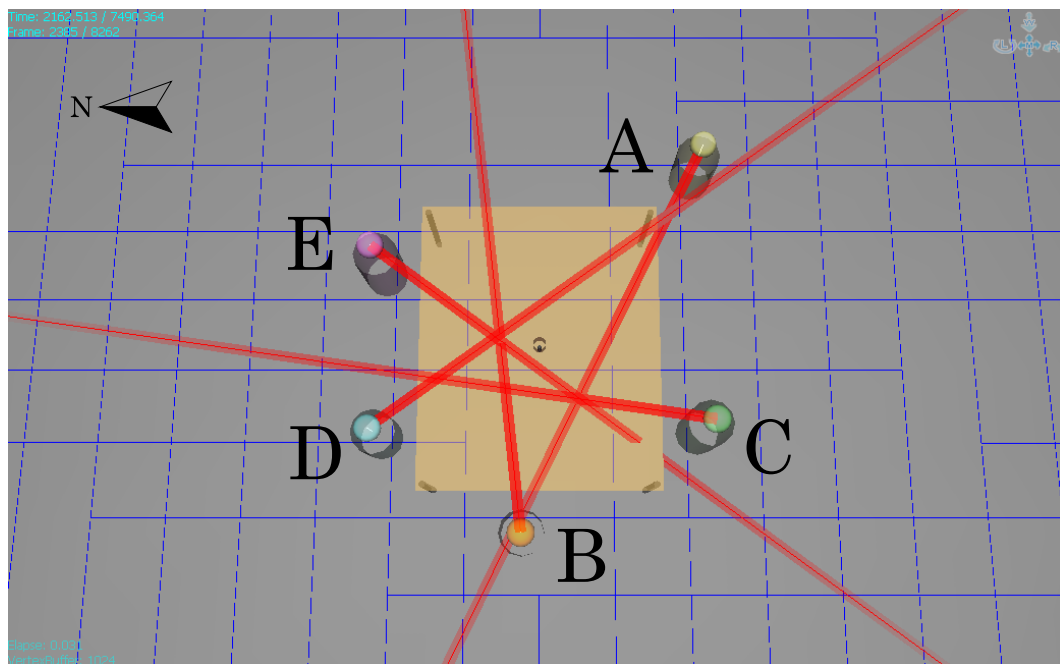


図 17: 参加者の配置

4.1 視線集中のインデクス

本システムでは顔方向と視線方向を検出できる．それぞれに対して集中の状況を計算し，その時刻周辺に現れる会話の性質について検証を行った．

4.1.1 顔方向集中

表 2 のようなインデクスが検出された．以下では，顔方向集中インデクス識別番号 1 を顔 index(1) などと表す．

顔方向集中のうち誤検出は，顔 index(1)，顔 index(2)，顔 index(3)，顔 index(4)，顔 index(5)，顔 index(11) であった．誤検出の原因としては以下のものであった．

- 顔方向は集中しているが視線は手元のコンピュータを見ている

表 2: 顔方向集中インデクス

識別番号	時刻	インデクス	備考
1	11:15:57	東面	誤検出
2	11:19:07	東面	誤検出
3	11:19:30	東面	誤検出
4	11:20:10	南面	誤検出
5	11:20:56	南面	誤検出
6	11:31:28	東面	方向誤り
7	11:31:58	東面	方向誤り
8	11:45:38	西面	方向誤り
9	11:48:38	北面	
10	11:48:55	北面	
11	11:51:42	北面	誤検出
12	12:00:34	北面	
13	12:03:32	東面	
14	12:07:10	北面	
15	12:07:13	南面	

- 顔追跡の失敗

また視線集中は起こっているがインデックスの方角が間違っていたものは、顔 index(6)，顔 index(7)，顔 index(8) であった。インデックスは東面となっているが、実際の視線集中は南面東側だったような場合がこれに該当する。原因は、例えば図 17 の B，C，D が同時に南面の A に注目するとき、同じく南面に属している C の顔方向はほぼ必ず東面を指すからである。

その他のインデックスは視線集中が起こっている面を正しく検出していた。視線集中が起こったときの会話内容の例を以下に示す。

顔 index(6) —

D「まあちょっと、しっかりデータを取るんだったら Mobile Eye のほうがいいんじゃないかなあと」

A「おー」

D「まあちょっと、金具さえ何とかなれば…」

A「結局こう、これだけ色々借りとして、Mobile Eye 買っちゃうなんていうことになったりして」← B,C,D,E 視線集中

一同「(笑い)」

顔 index(13) —

A「いわゆる DeCoPro 型のやつ…，あそこ(東側のホワイトボードを指差しながら)に書いてずっと残したままだけど、DeCoPro 型のやつで 2, 3 年間堂々巡りしているだけじゃん」

一同「(指差しにつられて東側に顔を向ける)」

顔 index(14), 顔 index(15) —

D「… ポスター上を指してっていうのはあるんですが、ここ(体の前で手をかざしながら)でやるようなジェスチャになるものは多分、何のジェスチャしてた？」← A,B,C 視線集中

C「一応なんか、地図の中にいるから、地図があるとして、こっち(ジェスチャをしながら)、とか、こういうのが…」← A,B,D,E 視線集中

視線集中が起こる会話場面の特征としては、以下のようなものであった。

- 意思決定に関わるような話題
- 指示代名詞とジェスチャを同時に用いたとき

- 質問に対する応答を期待して顔を向ける
- 比較的な大きな声で会話の主導権を取ったとき

例えば顔 index(6) では研究機材購入の意思決定の話題の中で，意外な結論に達しそうなため皆で笑い出すという場面である．また顔 index(13) や顔 index(14) では，指示代名詞とジェスチャをともに用いることで，参加者が見て確認せざるを得ない状況を作り出しているといえる．特に顔 index(13) では，図 18 のように机から離れた注目対象についても取り扱える可能性を示唆した．



図 18: 机から離れた注目対象への視線集中

4.1.2 視線方向集中

表 3 のようなインデックスが検出された．全 52 インデックス中，誤検出は 16 個，検出は 36 個である．方向誤りは 10 個であった．以下では，視線方向集中インデックス識別番号 1 を視 index(1) などと表す．

誤検出の原因としては以下のものであった．

- レンズ境界部分での顔のゆがみ
- まばたき

表 3: 視線方向集中インデクス

識別番号	時刻	インデクス	備考	識別番号	時刻	インデクス	備考
1	11:18:36	東面	方向誤り	27	11:49:29	東面	誤検出
2	11:19:52	南面	誤検出	28	11:50:14	南面	
3	11:20:24	東面	方向誤り	29	11:51:02	東面	方向誤り
4	11:20:47	東面	誤検出	30	11:53:59	東面	誤検出
5	11:21:11	南面		31	11:54:43	東面	方向誤り
6	11:21:15	東面	誤検出	32	11:55:28	北面	
7	11:23:49	北面	誤検出	33	11:56:22	北面	
8	11:26:22	東面	方向誤り	34	12:00:06	北面	
9	11:31:11	北面		35	12:00:12	北面	
10	11:34:55	北面		36	12:00:15	北面	
11	11:35:32	南面	誤検出	37	12:00:20	北面	
12	11:38:54	南面	誤検出	38	12:00:24	北面	
13	11:39:42	南面	誤検出	39	12:00:30	北面	
14	11:41:15	北面	誤検出	40	12:00:34	北面	
15	11:42:22	西面	方向誤り	41	12:01:04	北面	
16	11:42:23	南面		42	12:03:50	東面	
17	11:42:26	南面	誤検出	43	12:03:56	東面	
18	11:43:02	西面	方向誤り	44	12:04:43	東面	方向誤り
19	11:43:04	北面		45	12:06:39	北面	誤検出
20	11:43:30	北面	誤検出	46	12:06:42	東面	方向誤り
21	11:48:14	南面		47	12:06:43	南面	
22	11:48:25	北面	誤検出	48	12:07:00	北面	
23	11:48:37	北面		49	12:07:09	北面	
24	11:49:00	南面		50	12:56:05	東面	
25	11:49:07	南面		51	12:56:28	東面	誤検出
26	11:49:09	東面	方向誤り	52	13:00:15	東面	誤検出

- 解像度不足による眼球位置の取得不全

視線集中が起こったときの会話内容の例を以下に示す．

視 index(50)

A「え，月曜日なんかでもう何か話ができるの？」

D「いや，だからまあ金曜日が一番 …」

C「まあ，その明け一週間後で …」

A「金曜日？金曜日ちょっと僕出かけたい可能性が … あ，そうでもないや」

← B,C,D 視線集中

顔方向インデックスのほうが比較的白熱した部分に，視線方向集中インデックスのほうはより穏やかな部分に現れやすいようである．顔方向集中と共通するインデックスが見られる一方で，一人が話し続ける途中部分のインデックスが9個，質問応答の伴わない発話交代のインデックスが5個生成された．応答が返ってくるまで時間がかかる場面も散見された．これは，顔ごと向けて会話に参加する内部状態と，顔方向はそのまま視線方向だけ動かす参加の内部状態が異なるからではないかと考えられる．

4.2 実時間タグによるインデックス

実時間タグは，カメラ起動イベントが6，撮影イベントが13，書き込みイベントが47であった．生成された画像は撮影イベントと同じになり13枚である．カメラ起動イベント数と撮影イベント数が一致しないのは，カメラを起動したが写真をとらなかった場合と，カメラを起動せずに白紙を作成した場合が含まれるからである．13の撮影イベントのうち，カメラによる撮影が5，白紙が8である．カメラを起動したにもかかわらず，撮影を行わなかったイベントが1つあった．これは直後にカメラを起動し撮影を行っているため，カメラを構えなおしたか，他の写真を一度確認しなおした上で改めて撮影しなおしたと考えられる．

これらのイベントから，各画像の活動時間に注目する．活動時間の定義は，最初に写真撮影か白紙メモ作成によって書き込み領域が生成されてから，最後に書き込みが起こるまでである．この状況を表4に示す．この活動時間は話題の移り変わりと関係しているものと考えられる．すなわち活動時間が短いものは短期的な話題を，長いものは長期的な話題を包含していると考えられる．

図 19 に活動時間の長かった PhotoChat 画像を示す．

表 4: 画像ごとの活動時間

ID	種別	生成者	書き込み人数	書き込み回数	生成時刻	活動時間	備考
1	写真	A	1	1	11:10:12	0 分 6 秒	準備中
2	メモ	E	3	4	11:11:48	2 分 54 秒	落書き
3	写真	D	3	4	11:15:14	9 分 18 秒	全方位カメラ
4	メモ	B	1	7	11:16:20	33 分 31 秒	洛々とポスター
5	写真	A	3	4	11:24:45	1 分 7 秒	メール
6	メモ	B	3	5	11:26:35	20 分 2 秒	研究機材比較
7	メモ	E	4	8	11:35:45	14 分 50 秒	todo
8	メモ	A	2	4	11:35:55	23 分 27 秒	ポスター
9	メモ	B	1	1	11:57:49	0 分 3 秒	アイマーク水増し
10	写真	B	1	1	12:04:12	1 分 37 秒	ホワイトボード
11	メモ	B	2	3	12:04:18	1 分 47 秒	ホワイトボードの内容
12	メモ	E	1	3	12:19:44	1 分 56 秒	視線の分析
13	写真	E	1	2	12:27:41	8 分 45 秒	ホワイトボード
平均	-	-	2.0	3.6	-	9 分 11 秒	-

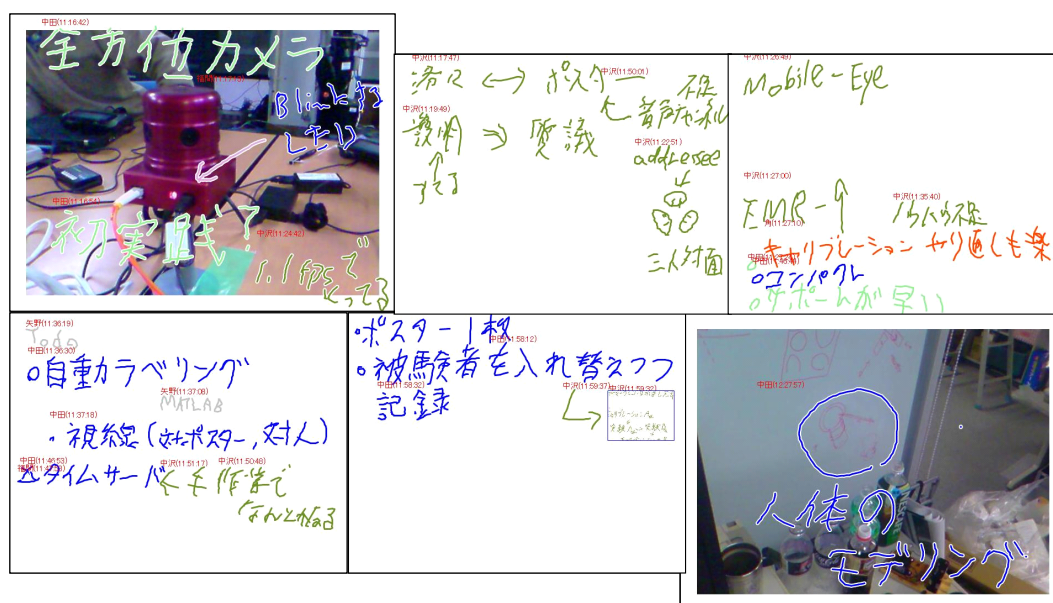


図 19: 活動時間の長かった PhotoChat 画像

4.3 実時間タグと視線集中の比較

実時間タグによるインデックスの時刻と顔方向集中インデックスの時間差を比較する．表 5 は顔方向集中インデックスを中心とした，最も近い実時間タグイベン

トとの時間差である．多くの場合 50 秒以上の時間差があり，これは視線集中という活動と，PhotoChat の操作という活動が両立しないことを示している．したがって参加者は会話の場への積極的な内部状態と，一段積極性を落とした PhotoChat への書き込みという内部状態を使い分けていると考えられる．

また顔方向集中に従属して実時間タグが発生する可能性を検討したが，時間差にその傾向は現れなかった．

表 5: 顔方向集中時刻から実時間タグまでの時間差

識別番号	負差	正差
6	-4 分 09 秒	4 分 05 秒
7	-4 分 39 秒	3 分 35 秒
8	-7 分 01 秒	1 分 02 秒
9	-0 分 49 秒	1 分 16 秒
10	-1 分 06 秒	0 分 59 秒
12	-1 分 09 秒	3 分 10 秒
13	-4 分 07 秒	0 分 12 秒
14	-1 分 02 秒	12 分 37 秒
15	-1 分 05 秒	12 分 34 秒

第5章 おわりに

5.1 結論

本研究では，プレゼンテーションやフリーディスカッションなどのミーティングのデータを再利用可能な状態にすることを目的として，ミーティングを収録する環境と，それらにインデクスを生成する方法論を検討した．収録はカメラアレイ，全方位カメラ，クラスタ計算機，PhotoChat，スライド操作検知を用いた．これらから得られるデータを基にインデクス生成を行った．

視線集中の判定は適合率 60.0%であり，視線集中の起きた時間帯の会話にいくつかの特徴的な性質を見つけることができた．実時間タグのインデクスでは活動時間に注目することにより，話題の継続性への手がかりとすることができる．これら二つのインデクスは互いに相補的な関係にあり，目的によって使い分ける必要がある．

5.2 今後の課題

本研究の手法により獲得した文字情報とインデクスを基に，ユーザに関係のありそうな過去の会話シーンのうち，自らが不参加だったものについて自動的に推薦を行ってくれたり，現在の会話の状況と関連するような過去の会話シーンを自動的に検索してくれたりするようなシステムの可能性を検討する必要がある．高品質なシーン映像を切り出すためには会話区間の判定が必要であり，時間制約と会話区間を用いることによって，シーン抽出が行える．例えば重要シーンインデクス前後 10 秒を上限として，発話交代が起こったところで切り分けるという方法が考えられる．

記録の再利用のことを考えたとき，ミーティング中に使用しているスライドを使うことが考えられる．その方法としては，リアルタイムでプレゼンテーションを行っているスライドを文書として扱い，類似文書検索の手法によって過去に行ったプレゼンテーションのスライドを抽出してくる．そのスライドが過去に用いられた時間のミーティング映像を推薦することで，再利用を促すことができる．

また視線方向のクラスタリングには単純なモデルを使用したため，視線集中の判定精度を上げるためには，さらにふさわしいモデルを構築することが課題となる．特に本稿による手法では自分が属している方向に対する注目をうまく

抽出できていない．これをうまく切り分けるには顔の方向と注目の方向を直線で定義するだけではなく新たな方法を検討する必要がある．

顔方向と視線方向の比較では，顔方向のほうが重要シーンインデクスを生成するためによりふさわしいことが示唆された．これは参加者の積極性の度合いがシーンに対して重要な意味合いを持つからであると考えられる．そこで今後，時刻の移り変わりに対する参加者の積極性を抽出していく必要がある．その方法としては，場の中央への体の乗り出し具合の変化や，背筋の伸ばし具合の変化を捉えることが考えられる．これらはカメラに対する顔の相対位置を検出することで，推定することが可能である．このような情報から重要シーンと積極性の関連を調べていく必要がある．

謝辞

本研究を進めるにあたって，多大なる御指導を賜った西田豊明教授，角康之准教授，ならびに貴重なご意見を賜りました角所考准教授に厚く御礼申し上げます．また，クラスタコンピュータ InTrigger を使うにあたって，田浦健次朗准教授，鴨志田良和助教をはじめとする東京大学 近山・田浦研究室の InTrigger コアメンバーには大変お世話になり，ここに改めて感謝の意を表したい．さらに顔センシングライブラリを使用するためのラッパーソフトウェアを提供くださった古川拓也氏，3次元空間上での計算・可視化を実現するソフトウェアを提供くださった矢野正治氏，PhotoChat デバイスからの時刻同期したログ抽出を実現するソフトウェアを提供くださった古谷翔氏，三氏のおかげで迅速な開発・分析が可能になったこと，大変感謝を申し上げたい．くわえて福間良平氏には本稿執筆中，多大なる励ましと助言をいただき，謝意に堪えない．そして，研究生生活を支援頂いた京都大学 西田・角研究室の皆様に深く感謝する次第である．

参考文献

- [1] Chiu, P., Kapuskar, A., Reitmeier, S. and Wilcox, L.: Room with a rear view: Meeting capture in a multimedia conference room, *IEEE Multimedia*, Vol. 7, No. 4, pp. 48–54 (2000).
- [2] Uchihashi, S., Foote, J., Girgensohn, A. and Boreczky, J.: Video Manga: Generating semantically meaningful video summaries, *Proceedings of Multimedia'99*, ACM, pp. 383–392 (1999).
- [3] Renals, S.: AMI: Augmented Multiparty Interaction, *Proc. NIST Meeting Transcription Workshop*, Montreal (2005). AMI-10.
- [4] Chen, L., Rose, R. T., Qiao, Y., Kimbara, I., Parrill, F., Welji, H., Han, T., Tu, J., Huang, Z., Harper, M., Quek, F., Xiong, Y., McNeill, D., Tuttle, R. and Huang, T.: VACE multimodal meeting corpus, *Second International Workshop on Machine Learning for Multimodal Interaction (MLMI 2005)*, Lecture Notes in Computer Science, Vol. 3869, Springer, pp. 40–51 (2006).
- [5] 友部博教, 土田貴裕, 大平茂輝, 長尾確: ディスカッションメディア: 会議コンテンツの構造化と効率的な閲覧システム, 第21回人工知能学会全国大会, 人工知能学会 (2007).
- [6] Abowd, G. D.: Classroom 2000: an experiment with the instrumentation of a living educational environment, *IBM Systems Journal*, Vol. 38, No. 4, pp. 508–530 (1999).
- [7] 西口敏司, 亀田能成, 角所考, 美濃導彦: 大学における実運用のための講義自動アーカイブシステムの開発, 電子情報通信学会, Vol. J88-D-II, No. 3, pp. 530–540 (2005).
- [8] Stiefelhagen, R., Yang, J. and Waibel, A.: Modeling focus of attention for meeting indexing, *ACM Multimedia '99*, ACM, pp. 3–10 (1999).
- [9] Otsuka, K., Araki, S., Ishizuka, K., Fujimoto, M., Heinrich, M. and Yamato, J.: A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization, *IMCI '08: Proceedings of the 10th international conference on Multimodal interfaces*, New York, NY, USA, ACM, pp. 257–264 (2008).
- [10] 久保田秀和, 齊藤憲, 角康之, 西田豊明: 会話量子化器を用いた会話場面の

- 記録, 情報処理学会論文誌, Vol. 48, No. 12, pp. 3703–3714 (2007).
- [11] 伊藤惇, 角康之, 久保田秀和, 西田豊明: Photo Chat : 互いの視点画像に「書き込む」ことによるコミュニケーション支援, 第 20 回人工知能学会全国大会, 人工知能学会 (2006).
 - [12] 角康之, 伊藤惇, 西田豊明: PhotoChat: 写真と書き込みの共有によるコミュニケーション支援システム, 情報処理学会論文誌, Vol. 49, No. 6, pp. 1993–2003 (2008).
 - [13] 斎藤秀雄, 鴨志田良和, 澤井省吾, 弘中健, 高橋慧, 関谷岳史, 頓楠, 柴田剛志, 横山大作, 田浦健次郎: InTrigger : 柔軟な構成変化を考慮した多拠点に渡る分散計算機環境 (HPC-14 : 分散処理), 情報処理学会研究報告. [ハイパフォーマンスコンピューティング], Vol. 2007, No. 80, pp. 237–242 (2007).
 - [14] 來嶋宏幸, 坊農真弓, 角康之, 西田豊明: マルチモーダルインタラクション分析のためのコーパス環境構築, 情報処理学会研究報告. ヒューマンコンピュータインタラクション研究会報告, Vol. 2007, No. 99, pp. 63–70 (2007).
 - [15] 瀧川えりな, 細井聖, 川出雅人: 話題の製品と技術性別、年代、人種を見分ける—顔属性推定技術「OKAO Vision」, 画像ラボ, Vol. 15, No. 4, pp. 82–87 (2004).
 - [16] Bradski, G. R.: Computer Vision Face Tracking For Use in a Perceptual User Interface, *Intel Technology Journal*, No. Q2, p. 15 (1998).
 - [17] 工藤拓: MeCab: Yet Another Part-of-Speech and Morphological Analyzer (2006).
<http://mecab.sourceforge.net/>.
 - [18] Huynh, D. F.: SIMILE Timeline. <http://simile.mit.edu/timeline/>.