MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI230972

Individual Activity Anomaly Estimation in Operating Rooms Based on Time-Sequential Prediction

Koji YOKOYAMA [a,1], Goshiro YAMAMOTO [a], Chang LIU [a], Kazumasa KISHIMOTO [a], Yukiko MORI [a] and Tomohiro KURODA [a]
[a] *Kyoto University*

Abstract. Surveillance videos of operating rooms have potential to benefit post-operative analysis and study. However, there is currently no effective method to extract useful information from the long and massive videos. As a step towards tackling this issue, we propose a novel method to recognize and evaluate individual activities using an anomaly estimation model based on time-sequential prediction. We verified the effectiveness of our method by comparing two time-sequential features: individual bounding boxes and body key points. Experiment results using actual surgery videos show that the bounding boxes are suitable for predicting and detecting regional movements, while the anomaly scores using key points can hardly be used to detect activities. As future work, we will be proceeding with extending our activity prediction for detecting unexpected and urgent events.

Keywords. Video Analysis, Operating Room, Individual Activity, Anomaly Estimation, Time-Sequential Prediction

1. Introduction

Recently, surveillance cameras are being installed into operating rooms. Such surveillance videos of operating rooms have potentials to benefit specific post-operative analysis and study, e.g. teamwork evaluation and emergency management analysis. However, there is currently no effective method of extracting useful information about surgical activity from the records without human intervention. In addition, as a surgery with the preparation phase usually continues for multiple hours, it is inefficient to analyze the intraoperative videos manually [4].

As a first attempt to this issue, we study about the effective methods to automatically extract individual activities from operating room surveillance videos. We propose a semi-supervised individual activity anomaly estimation model based on time-sequential prediction using Generative Adversarial Network (GAN) [3]. As it is difficult to label intraoperative activities, we take unsupervised model. In this paper, we compare two specific features that can be used as inputs to our method to acquire anomaly scores.

[1] Corresponding Author: Koji YOKOYAMA, Ph.D Student, Department of Social Informatics, Graduate School of Informatics, Kyoto University, Japan, Tel: +81-075-336-7703; E-mail: yokoyamak@kuhp.kyoto-u.ac.jp.

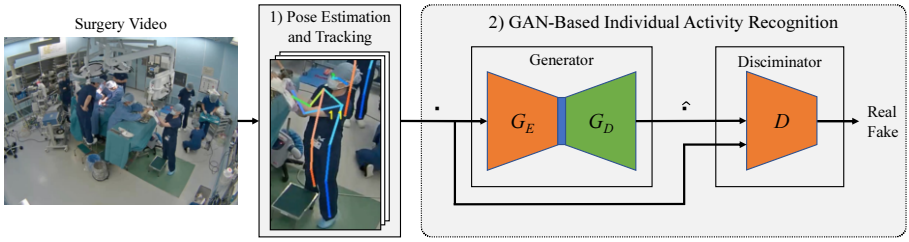

Figure 1 System flow comprising two components: 1) pose estimation and tracking, and 2) GAN-based individual activity recognition.

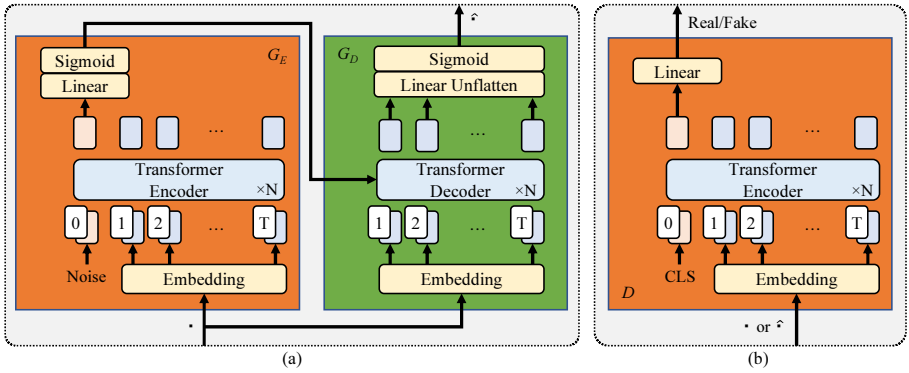

Figure 2 The architecture of GAN-based individual activity recognition model which consists of (a) generator and (b) discriminator.

2. Methods

This section introduces the details of our method. Figure 1 shows system flow of proposed method which is composed of two major components: 1) pose estimation and tracking, and 2) activity anomaly estimation using time-sequential prediction.

2.1. Pose Estimation and Tracking

Firstly, we extract bounding boxes and key points of each individual from surgery videos as shown in Figure 1. We used YOLOX [2] for extracting bounding box, High-Resolution Net (HRNet) [6] for pose estimation, and COCO [5] for the key points which includes 5 face points and 12 body points. We also used UniTrack [9] to track individuals' key points, the results are shown by yellow color numbers.

2.2. Activity Anomaly Estimation using Time-Sequential Prediction

We define individual activity anomaly as difference between real data and fake data generated by GAN. Figure 2 shows the architecture of our individual activity anomaly estimation model which is similar to skip-GANomaly [1]. We customized the architecture to use the transformer layer [7] instead of a convolutional neural network.

The model is fed with time-sequential input features x into the model. The dimension of x is $(t, n, 2)$, where t is the number of frame and n is the number of points.

All points are calculated in two-dimensional coordinates. The generator consisted of a transformer encoder G_E and a transformer decoder G_D, which is able to predict fake activity data $\hat{x}$. The discriminator D consisted of a transformer encoder, which catches the input x or $\hat{x}$ and predicts if the input is real or fake. G_E and D have internal parameters that are noise parameters (Noise) and class token (CLS), respectively. All of them have same embedding layers and positional embedding layers. The embedding layer consists of two pointwise convolution layers. The first one convolutes the coordinates of x to one-channel value, while the second one extracts distributed representation of x in each frame t. Positional embedding layer uses learnable absolute positional encoding [8]. Finally, we estimate individual activity anomaly scores using the mean squared error of x and $\hat{x}$.

3. Results

3.1. Dataset

Out of surveillance videos (30 FPS) from 6 operating rooms, we selected and cropped 40 pieces of short videos of which each long for 3 minutes. Half were used as training datasets where the surgeries proceed smoothly, and the other half were test datasets including pre- and post-operative phases, and temporary suspending.

We fed two types of 10 seconds ($t = 300$ frames) of x into our model. The first one was bounding box that was scaled by dividing by the frame size of the surveillance video. The second one was key points substituted by the top-left corner of the bounding box and scaled by dividing by the size of bounding box. We trained our model on each dataset.

3.2. Implementation

We trained the models while 100 epochs with batch size of 4096. The contextual-loss was changed to Mean squared Error from skip-GANomaly [1]. The optimizer for the learning model was Adam. The learning rate was $lr = 0.001$, and betas were $(\beta_1, \beta_2) = (0.5, 0.999)$.

3.3. Results

The anomaly scores are shown in Figure 3: (a) one of training datasets which shows proceeding smoothly, and (b) one of test datasets which shows temporary suspending.

For the top of Figure 3 (a), there were two peaks on the orange line (tracking ID: 272), in which several activities occurred such as walking, throwing away trash, crouching, and stepping up. Figure 4 (a) shows the results of x and $\hat{x}$ indicated by the red arrow of Figure 3 (a). In this situation, the circulator did two activities: (1) changing the position of the stage while crouching, and (2) stepping up on the stage. Then, the bounding box results indicates there were large errors between real bounding boxes and fake bounding boxes. However, for the bottom of Figure 3 (a), all of the lines are under 0.65. The bottom of Figure 4 (a) indicates there were little errors between real key points and fake key points.

For the top of Figure 3 (b), anomaly scores of the blue line (tracking ID: 34), the orange line (tracking ID: 39), and the green line (tracking ID: 56), exceeded 0.05. In

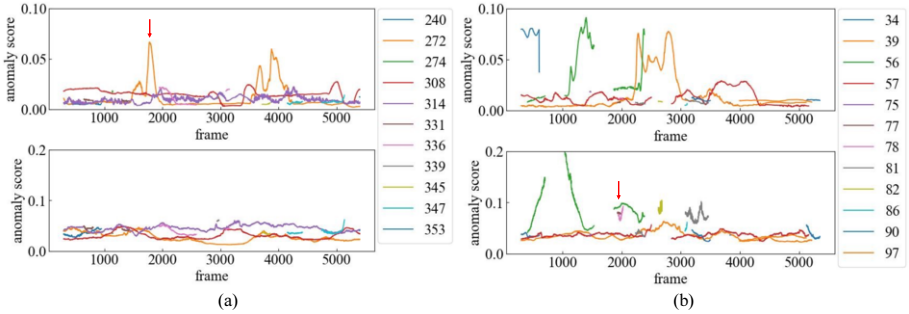

Figure 3 The result samples of anomaly score: (a) training dataset and (b) test dataset. The top of each result was estimated by bounding boxes. The bottom of each result was estimated by key points. The labels are tracking IDs. The red arrows show the situation of Figure 4 (a) and (b).

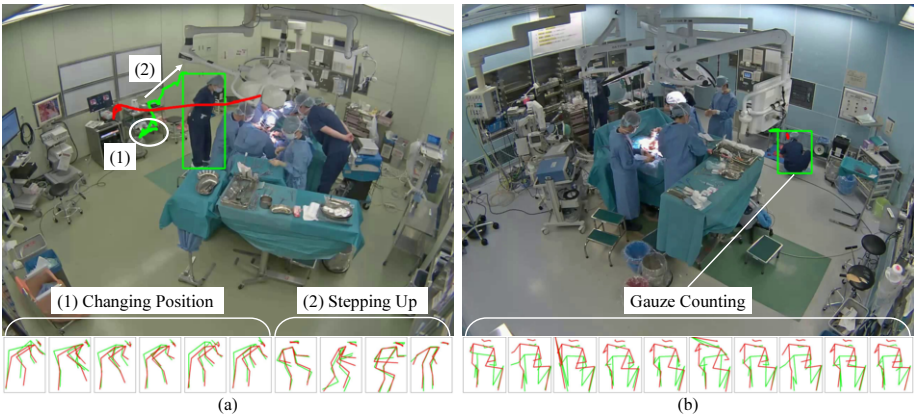

Figure 4 The real data and fake data indicated by the red arrows in Figure 3: (a) training datasets and (b) test datasets. Green and red lines show the real data and the fake data, respectively. The top of each shows the top-left corner of bounding box and the bottom of each shows key points at 30 frames (1 sec) intervals.

addition, for the bottom of Figure 3 (b), anomaly scores of the green line (tracking ID: 56), the yellow line (tracking ID: 82), and the gray line (tracking ID: 81) exceeded 0.1. Several activities also occurred at the point of high anomaly score such as walking, crouching and medical instrument passing. Figure 4 (b) shows the red arrow of Figure 3 (b) in which the circulator counted gauze while crouching. For Figure 3 (b), as she remained same position for 300 frames, little errors the bounding box results were occurred. However, large errors of the key points results were occurred.

4. Discussion

When the anomaly scores from bounding boxes were observed high, large movement activities occurred such as walking around. When surgery was proceeding smoothly, the surgical members did hardly move except for the circulator. Figure 3 (a) indicates that the surgery proceeded smoothly because the anomaly score of the circulator (tracking ID: 272) was high. As shown in Figure 3 (b), there were several high anomaly scores because surgical members temporarily went away from the operating table. It is indicated that the prediction using bounding boxes is suitable for detecting and estimating regional movements of the individuals.

Based on the bottom of Figure 3 (b), we looked into the videos and found some error peaks in the anomaly scores extracted from key points. It is caused by occlusion in the videos, except for some cases where the members were clear to the camera or avoiding the occlusion occasionally, e.g. the member crouching and counting gauze in Figure 4. Therefore, the prediction using key points could hardly be used to detect unique activities in the surgery. We will look into masking low confidence key points to solve this issue.

In this paper, we utilized cropped videos which do not include any irregular events. In future, we will extend and improve our activity prediction to fit for detecting unexpected and urgent events by applying our model to videos including irregular issues and analyzing the anomaly scores with other latent features.

5. Conclusions

We investigated automatic analysis of intraoperative activities from surveillance videos of operating rooms. We proposed a novel method to automatically estimate individual activities using anomaly estimation based on time-sequential prediction.

We evaluated the performance of the model with two types of time-sequential features: bounding boxes and key points. Experiment results show that the bounding boxes are suitable for predicting and detecting regional movements, while the anomaly scores using key points can hardly be used to detect activities. In future, we will consider key event detection by applying our method to the beginning to the end of surgery videos.

Acknowledgements

This work was partially supported by JSPS KAKENHI Grant Numbers JP22H03632.

References

[1] S Akcay, A Atapour-Abarghouei, TP. Breckon. Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection. In *2019 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE, 2019.

[2] Z Ge, S Liu, F Wang, Z Li, J Sun. YOLOX: Exceeding yolo series in 2021, *arXiv preprint arXiv:2107.08430*, 2021.

[3] IJ. Goodfellow, J Pouget-Abadie, M Mirza, B Xu, D Warde-Farley, S Ozair, A Courville, Y Bengio. Generative adversarial networks, *arXiv preprint arXiv:1406.2661*, 2014.

[4] Costa Jr, A da Silva. Assessment of operative times of multiple surgical specialties in a public university hospital. Einstein, 15:200–205, 2017.

[5] TY Lin, M Maire, S Belongie, J Hays, P Perona, D Ramanan, P Doll´ar, C.L Zitnic. Microsoft COCO: Common Objects in Context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T(eds) Computer Vision – ECCV 2014. Lecture Notes in Computer Science, vol 8693. Springer, Cham.

[6] K Sun, B Xiao, D Liu, J Wang. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5693–5703, 2019.

[7] A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, AN. Gomez, L Kaiser, I Polosukhin. Attention is all you need, *arXiv preprint arXiv:1706.03762*, 2017.

[8] B Wang, L Shang, C Lioma, X Jiang, H Yang, Q Liu, JG Simonsen. On position embeddings in BERT. In *International Conference on Learning Representations*, 2021.

[9] Z Wang, H Zhao, Y Li, S Wang, P Torr, L Bertinetto. Do different tracking tasks require different appearance models? In M Ranzato, A Beygelzimer, Y Dauphin, P.S Liang, J Wortman Vaughan, editors, Advances in Neural Information Processing Systems, 34:726–738. Curran Associates, Inc., 2021