A Group Activity Based Method for Early Recognition of Surgical Processes Using the Camera Observing Surgeries in an Operating Room and Spatio-Temporal Graph Based Deep Learning Model

Keishi Nishikawa[1] and Jun Ohya[2]
[1]*Global Information and Telecommunication Institute, Waseda University, 3-4-1, Okubo, Shinjuku, Japan*
[2]*Department of Modern Mechanical Engineering, Waseda University, 3-4-1, Okubo, Shinjuku, Japan*

Keywords: Graph Processing, Early Recognition, Surgical Process, Operating Room.

Abstract: Towards the realization of the scrub-nurse robot, this paper proposes a group activity-based method for the early recognition of surgical processes using an early part of the video acquired by the camera observing surgeries. Our proposed method consists of two steps. In the first step, we construct a spatial-temporal graphs which represents the group activity in operating room. The graph's node contains (a) the visual features of participants and (b) the positions. In the second step, the generated graphs are input to our model for classification of the input. In our model, since the generated graph' node contains both visual features and the position information, we treat the graph as the point cloud in spatial-temporal space. Therefore, Point Transformer Layer from (Zhao et al., 2021) is used as the building block. Experiments are conducted on public datasets; (Özsoy et al., 2022)'s mock surgery of knee replacement. The results show our method performs early recognition achieving the accuracy of 68.2 %~90.0 % in early duration such as 17.1 % ~ 34.1 % of the entire durations from the beginning on the dataset. Furthermore, the comparison with the state-of-the-art method (Zhai et al., 2023) in early recognition of group activity is also conducted. It turns out that ours outperforms (Zhai et al., 2023) significantly.

1 INTRODUCTION

In general, a surgery is conducted with the cooperation of both many surgeons and nurses. Towards keeping the surgery forward smoothly, the nursers often observe the situation, recognize the necessary instruments and information, and pass them to the surgeons. However, according to the investigation by The Japan Institute for Labour Policy and Training, the shortage of the nurses is a serious problem (The Japan Institute for Labour Policy and Training, 2022). To relax this problem, it is necessary to achieve systems that work as alternatives of the nurses. (Li et al., 2016) indicate that the requirements of such a system include to recognize and make the appropriate decisions for supporting the surgeons autonomously, without any external operation by the surgeons, being able to recognize the surgical process and the actions as early as possible to pass the necessary instruments to the surgeons, and so on.

As shown in Fig. 1, the surgery consists of multiple surgical processes, and each of the surgical processes consists of one or more surgical actions, forming a kind of hierarchical relationship among the surgical processes and actions. The level shown in Fig.1 means the granularity in the hierarchy. Note that the process is at the higher level than the action. In (Li et al., 2016), the surgical actions to be recognized have been focused on the hand actions which relates to the requirement of passing the instrument to the surgeons towards replacing the nurses. However, it is also necessary to do the following things in addition to passing surgical tools.

1) Providing useful information about the cautions and the roles of the participants to the surgery.

2) Recognizing the emergencies that occurred during the surgical process and controlling the robots to help the surgeons avoid the difficult situation.

For achieving the above points, it is necessary to recognize the surgical process. In the view of providing the support for the surgery, each surgical process and action must be recognized as early as possible. Even though each surgical process consists

Nishikawa, K. and Ohya, J.
A Group Activity Based Method for Early Recognition of Surgical Processes Using the Camera Observing Surgeries in an Operating Room and Spatio-Temporal Graph Based Deep Learning Model.
DOI: 10.5220/0013252700003905

In *Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2025)*, pages 712-724
ISBN: 978-989-758-730-6; ISSN: 2184-4313

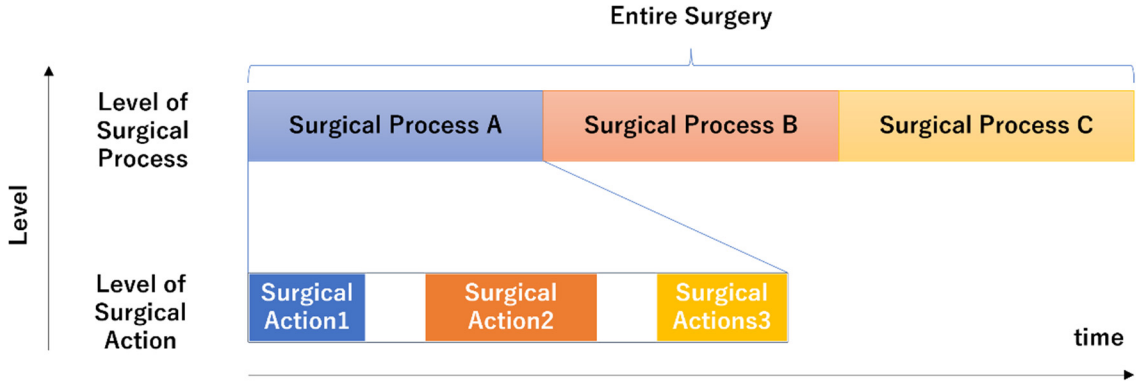


Figure 1: Flow of the surgery and the levels of the elements consisting of the surgery.

of multiple surgical actions, it is highly possible that multiple surgical processes could contain same surgical actions. Therefore, recognizing surgical actions is not sufficient for recognizing surgical processes.

One of the important features of the surgery is that the surgeons and the nurses could cooperate and interact with each other. Therefore, although the work (Li et. al, 2016) recognizes individual's actions, that is not sufficient for recognizing the surgical process, as mentioned earlier. To solve this issue, the motions , the positions, and the number of the participants in the operating room can be useful. Therefore, it is reasonable that the positions and motions of the participants as groups at each time instance are captured and used for the early recognition of surgical processes. For collecting that kind of information, a bird-view camera which can capture the entire space of the operating room is appropriate.

While some works about surgical process recognition have been conducted in the form of surgical phase classification or surgical activity recognition using videos acquired by the camera installed in an operating room (K. Yokoyama et al. 2023), (E. Özsoy et al. 2024), (E. Özsoy et al. 2023), (Shargi et al., 2020), (L. Bastian et al., 2022), these did not focus on early recognition. Therefore, to best our knowledge, early recognition of activities in an operating room has not been researched well. Furthermore, dataset collected in the community of surgical data science tends to be small compared to that in the community of computer vision due to privacy issue etc.

In this paper, a method for early recognition of the surgical process using a bird-view camera in the operating room is proposed. The proposed method consists of the following two modules: (1) constructing the spatial-temporal graphs which represents not only visual features of the participants but also the positions and the number of participants of the surgery and (2) processing the graph generated during early duration of the input video to classify it as one of the surgical processes by our model which consists of graph neural networks layer. Especially the model utilizes Point Transformer Layer as the building blocks to deal with the graphs which has the geometric information in spatial-temporal space. For conducting experiments, we used publicly released video dataset which observes mock surgery of knee-replacement.

2 RELATE WORK

2.1 Conventional Works

In case of action recognition, the goal is to recognize the action using the video data captured from the beginning to the end. However, in case of early recognition, the input to the recognition model is an early part of the video data: from the beginning to an early time instance before the end. According to the survey (H Zhao et al., 2021), the conventional works can be classified into the three main categories: (1) One-Shot based method, (2) Knowledge Distillation based method, and (3) Propagation based method.

2.1.1 One-Shot Based Method

The methods in this category do not take the fully observed information as the input, but the partially observed information which means the early part of the video as the input (Zhou et al, 2018), (Chen et al., 2018), (Singh et al., 2017), (Sun et al., 2019). (Zhou et al, 2018) exploit the features extracted from the entire area of the image. (Chen et al., 2018) focus on the joints of a person in the given video and weight the features of the local region which surrounds the

joints for improving the recognition of the action in the early stage. The method proposed by (Singh et al., 2017) detects the rectangle which surrounds the person in the video and achieve the early action recognition in real time setting by accumulating the detections over the temporal domain. (Sun et al., 2019) also detect the multiple people in the image and create the graphs whose node contains the feature, which is extracted from the detections: the graph is used for representing the relationship among the multiple people in the image.

2.1.2 Knowledge Distillation Based Method

Although only the partial observed information is used for the inference, both the full and partial information are used for the training. In this setting, the methods proposed in (Ma et al., 2016), (Aliakbarian et al., 2017) construct the model for increasing the confidence of the recognition as the observed information increases by the temporal progress. In (Aliakbarian et al., 2017) (Kong et al, 2017), the mapping function is learned by approximating the fully observed information with the partial information for improving the recognition at the early stage of the action.

2.1.3 Propagation Based Method

The methods in this category use the generative model for generating the data which will be likely to be observed in the future based on the partially observed information. One of the representative works in this category is the work by (Zhao et al., 2019).

2.1.4 Early Recognition of Group Activity

While many of the methods in the above categories are targeting at actions done by one person or targeting at scenes in which the actions are not treated as group actions, there is a work for recognizing group actions at the early stage.

(Chen et al., 2020) exploits two graphs for the input. The first graph represents the positions of the people and the relationship among the people. The second graph represents the features of the action of each person and the relationship based on the similarities among the actions. They use two types of autoencoders for extracting the features from both graphs respectively. The encoders are learned with the adversarial loss to approximate the fully observed information with the partially observed information. Chen et al. validate the effectiveness of their method for the early recognition of group actions in the volleyball and daily actions such as walking. They also note the predicting the positions is important to capture the group action.

(Zhai et al., 2023) also propose a method for early recognition of group actions. In (Zhai et al., 2023), a virtual leader node which connects to each individual node in the group is added to the graph while Chen et al. (Chen et al., 2020) focus on the pair-wise connection among the group. By setting virtual leader node, (Zhai et al., 2023) can summarize the group-wise representation. Also, (Zhai et al., 2023) do not require fully observed information for learning as opposed to (Chen et al., 2020).

Based on the above survey and our purpose, it is natural to consider that our method belongs to the works which are introduced in this section, that is early recognition of group activity. While there are many works about the recognition of group activity which recognize the activity based on the fully observed information (Wu et al, 2019), (Li et al., 2021), not many works on early recognition of group activity can be seen: to best our knowledge, (Chen et al., 2020) and (Zhai et al., 2023) are the only works. However, (Chen et al., 2020) and (Zhai et al., 2023) are targeting at general group activities, as opposed to our target at group activities in operating rooms.

2.1.5 Activity Recognition in Operating Room and Dataset

Conventional works about action/activity recognition (not early recognition) in an operating room have been conducted (Yokoyama et al., 2023), (Özsoy et al., 2024), (Shargi et al., 2020), (Bastian et al., 2022). (Yokoyama et al., 2023) proposed a method for the detection of passing a medical instrument and group attention during surgery by utilizing pose estimation of the participants. Experiments are conducted by using six videos capturing real surgeries by a camera attached to the ceiling of the operating room.

(Shargi et al., 2020) proposed a method for surgical activity recognition in an operating room in which they assume robot-assisted surgery would be conducted. They took video clips as the input from which features are extracted by a deep learning model and correlated the features to temporal contexts. They collected 400 videos captured by multiple time of flight cameras attached to the operating room. The size of their dataset could be the largest in the community of surgical data science, but the dataset is not released publicly.

(Özsoy et al., 2022) and (Özsoy et al., 2024) released the dataset which captures the simulated surgery of total knee replacement by multi-view

RGB-D cameras in an operating room. The dataset is called 4D-OR, which is the first publicly released dataset capturing the surgery in OR (an operating room) according to (Özsoy et al., 2022). In (Özsoy et al., 2022) they proposed a method for generating semantic scene graphs in the operating room and utilizing the graphs for predicting the roles of the participants in the surgery. They captured 10 surgeries for the dataset. Moreover, in (Özsoy et al., 2024) they extended their work (Özsoy et al., 2022) by adding the surgical phase recognition to the downstream tasks.

(Bastian et al., 2022) investigated the modalities of the input to the deep learning model for surgical action classification and which method for fusing them makes the best performance. In (Bastian et al., 2022), the modalities are RGB, depth and infrared because they use Asure Kinect Camera (Microsoft, 2021) which can capture RGB, Depth and Infrared images. They collected 16 videos of laparoscopic interventions in multi-view setting. They found that combination of RGB and Depth and late fusion of RGB and Depth made the best performance for surgical action classification.

To summarize, although the action/activity recognition in an operating room to which the camera is attached and capturing the surgeries in bird-view or objective-view are researched frequently, early recognition of group activity in an operating room is not conducted in the community of surgical data science. In addition to that, it is obvious that the dataset collected in the operating room towards the research in the community is much smaller than that in the community of the computer vision except (Shargi et al., 2020). Although the dataset collected in (Shargi et al., 2020) is much larger than those in the others, it is not publicly available.

2.2 Problems in Conventional Works

In this paper, the goal is the early recognition of the surgical process. Based on our observation, the two characteristic points of the surgical process are found.

(A) The surgical process is a lengthy operation lasting from tens of minutes to several hours.

(B) The surgical process is carried out simultaneously by multiple participants, including the surgeons, assistants, and nurses. Since each participant works at his/her own position in the operating room to play his/her own role. The number of the participants dynamically changes depending on the surgical process. Not only the visual features of the participants but also the position and the number for those can be distinguishable ques of activity in early stage.

In terms of the above-mentioned (A) and (B), the conventional works have the following problems.

In general, they used the datasets (Soomro et al., 2012), (Jhuang et. al, 2013), (Kong et al, 2012), (Patron-Perez et. al, 2010), (Goyal et. al, 2017), (Hu et al., 2017), (Y. Li , C. Lan and et. al, 2016), (Liu et al., 2017), which comprise short actions. Furthermore, both the knowledge distillation-based method and the propagation-based method use not only the partial information but also the fully observed information. In addition, the surgery takes a few hours: i.e. the duration of the surgery is much longer than that of the actions in the above-mentioned datasets. Therefore, the amount of the features increases as the observed frames increases. Though LSTM (Long Short Term Memory), which is often used, processes the temporal dynamics of the features (Ma et al., 2016), (Aliakbarian et al., 2017), it is difficult to build a realistic model that deals with temporally much longer data in a reasonable fashion. The reason is the low efficacy of the learning as pointed out in (Bradbury et al., 2016), while the learning in parallel is difficult.

The methods without using LSTM (Kong et al, 2017) and (Kong et al., 2014) divide the full-observed information into multiple partial information in the temporal dimension. After the division, multiple early image data are stacked to the matrix. In the learning, the mapping function is learned by approximating the stacked early image data matrix to the full-information matrix. This processing is affected by the temporal width (duration) of each division. In case of the smaller width, the temporal dimension of the stacked information gets larger because the amount of the features from the set of the partial information increases. That causes the computation during the learning to be difficult. In case of the larger width, the division gets coarser. This leads to the deterioration of the capability of the features to represent the action because the one feature corresponds to the early but long duration, which means that the information of the long duration is compressed into the one vector which constructs the early image data matrix.

One shot based methods use the short-duration information to recognize the actions. The input is the features extracted from the entire region (Zhou et al, 2018) or from the local region which surrounds the joints of the person (Chen et al., 2018), while those methods do not utilize the positions of people explicitly for building their models. In the videos capturing multiple persons' actions, (Singh et al.,

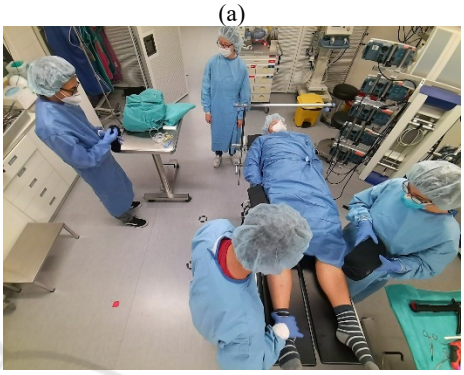
2017) detected the area of each person and represented the graph by connecting the detections. However, the node of the graph is the feature extracted from the areas of the detections and they focus on the relationship. Therefore, Singh et al. (Singh et al., 2017) do not utilize the positions and the motions of the people in the video either.

Although the camera in the operating room captures the entire area of the room, these methods cannot deal with the recognition capturing the characteristics described as the above-mentioned feature (B). The reason is that they extract the features from the entire region of the image, however, they do not embed the contribution of each person for group activity in the image explicitly.

In (Chen et al., 2020), which is a major work and similar one to ours in terms of early recognition of group activities, (Chen et al., 2020) represented the actions of the group using a graph-structure. They used multiple auto-encoders that are learned based on the partial information that is represented by a spatio-temporal graph. As mentioned above, longer videos tend to cause the problem which is similar to the one of (Kong et al, 2017) and (Kong et al., 2014). In addition to that, (Chen et al., 2020) uses both full and partial information for computing the adversarial loss. This also makes the computation difficult. In (Chen et al., 2020), their targets are short-duration-actions of groups such as spikes in volleyball (Ibrahim et al, 2016) and daily activities (Choi et al., 2009), including walking. The durations of these actions are about a few seconds, but, on the other hand, the dataset (Özsoy et al., 2022) which simulates the surgery contains processes whose durations are more than 1 minutes.

(Zhai et al., 2023) do not need the fully observed information for building the model and achieve state of the art performance on the datasets introduced by (Ibrahim et al, 2016) and (Choi et al., 2009). (Zhai et al., 2023) introduced a virtual leader node into a graph which represents the group activity. The node represents the summarization of the features of member's action by gathering visual feature vector from member's area in each frame. Furthermore, (Zhai et al., 2023) introduced Group Residual module which exploits the idea of virtual leader node for processing the graph. As a result, (Zhai et al., 2023) outperform (Chen et al., 2020) using the dataset of (Ibrahim et al, 2016) and (Choi et al., 2009) because (Zhai et al., 2023)'s method successfully captures characteristic points of group activity.

(a)



(b)

Figure 2: Examples of the images from surgical processes in (Özsoy et al., 2022).

However, (Zhai et al., 2023) did not consider the position and the number of the participants in each frame and balance among them. In this paper, note that members correspond to participants of surgeries. In the operating room, the participants have their own roles needed for accomplishing the surgical process, and the positions, the number and those variation are different from each other. This problem belongs to (B) because members' contribution and the temporal variation relates to the role of the members. In this paper, the method is proposed for early recognition to solve the problems (A) and (B): that is, the proposed method utilizes the motions, the potions and the number of the participants in the operating room.

3 PROPOSED METHOD

3.1 Our Basic Idea

As a result of our observations to the mock surgery (Özsoy et al., 2022), the following two points are found; (1) early recognition of surgical process in operating room is that of group activity in operating room, and (2) cues for early recognition are not only the visual features of the participants but also the

positions and the number of the participants. It is natural to assume that the visual features represent how the participants do their own task during the surgical process. In addition to that, the positions and number vary among the surgical processes. The examples are shown in Fig. 2. In Fig 2 (a), two participants are doing the task close to the table and the other participant is watching the scene at a place away from the two. On the other hand, in Fig 2 (b) there are four participants, and their positions are different from those in (a).

To reflect the visual features, the positions and the number of the participants for early recognition, the graph whose node represents the visual feature and the position of a participant is built at each frame and it is accumulated along the temporal progress so that the spatial-temporal graph is acquired. The number of participants is represented by the number of nodes in the graph at each frame. Here, from the point of view in geometric shape, the spatial-temporal graph can be regarded as point cloud data with other modal information such as visual feature at each point because the node has the position information therefore it can be viewed as point in space. Furthermore, the variation of the number, the position of the node and the temporal variation in the spatial-temporal graph can be the shape of point cloud data in spatial-temporal space. Based on the idea, a model which is used by our method is built based on graph neural networks which can process the input graph considering the property of the point cloud.

While our method is similar to (Chen et al., 2020) in terms of using spatial-temporal graph including the visual features and the 2D positions of the participants, our method is different from (Chen et al., 2020) because the graph contains the visual features, the positions of the participants in each frame and the time instance which corresponds to each frame in the form of the point cloud. In addition to that, ours differs from (Chen et al., 2020) in processing the graph because (Chen et al., 2020) treats the visual features and position information separately and process the two kinds of auto-encoders respectively, while ours processes our spatial-temporal graphs by utilizing the graph neural network which can deal with our graph as point cloud.

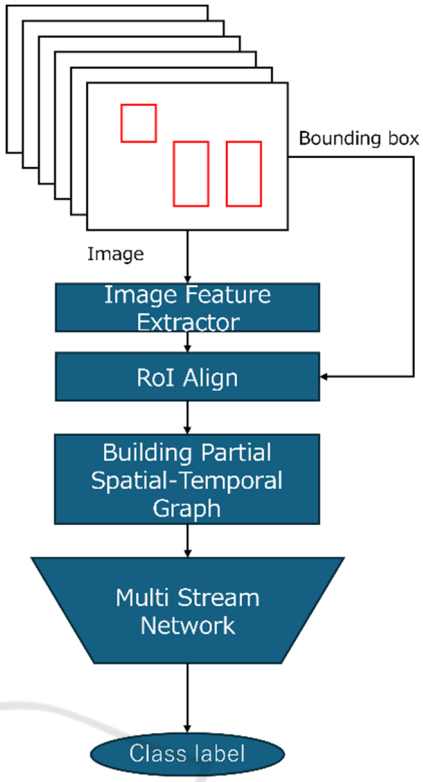


Figure 3: The entire flow of our proposed method.

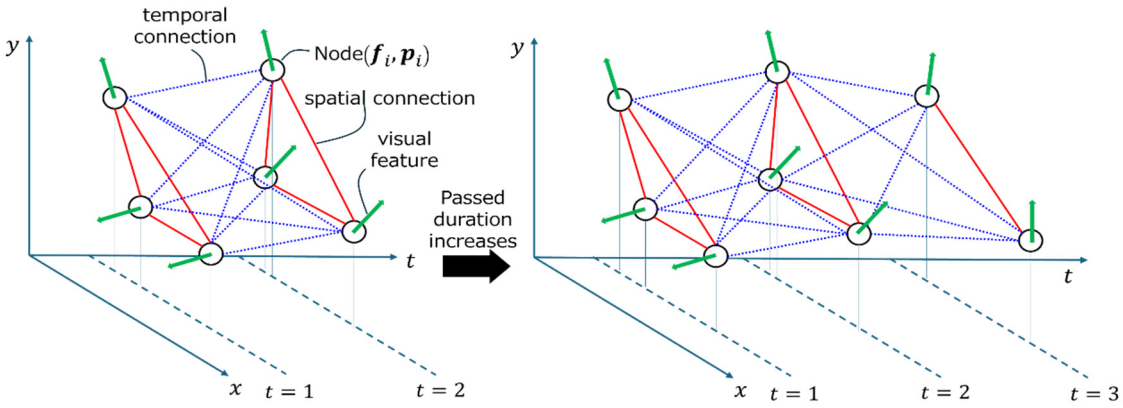


Figure 4: Overview of Generation of Graph.

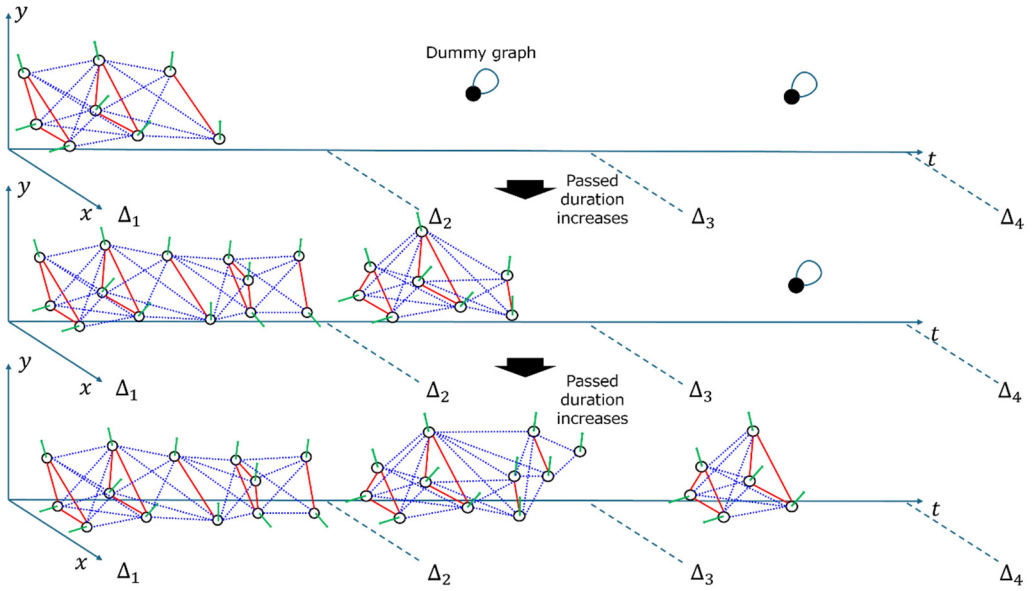


Figure 5: Generation of Patial Graph in which the number of temporal segments is three.

3.2 Flow of Our Proposed Method

The flow of our proposed method is shown in Fig.3. The proposed method consists of two steps. In the first step, the Spatial-temporal graphs are built. The graph represents the activity of the group during the partial duration. After building the graphs, these graphs are fed into a model. The model infers the class which the input graphs belong to. That is, we formulated the problem of early recognition of group activity in operating room as the classification of the spatial-temporal graph which is generated from video frames observed in early part.

3.3 Graph Construction

In this section, the process for making these graphs is explained. Note that it is assumed that the detections of the participants in each frame are obtained in advance.

3.3.1 Basic Idea of Generating Graph

The procedure for making the graph from the observed video frames is shown in Fig 4.

In the first step, the visual feature is extracted from the local area corresponding to the bounding box which surrounds each participant in the feature map of each frame by using RoI-Align(He et al., 2017). At the same time, the centre coordinates of the bounding boxes are also recorded as the geometric information.

In the second step, the visual feature and geometric information of each person are associated with a node of the graph. Note that, in this step, the time instance is added to the geometric information so that the geometric information has the 3D information which consists of the 2D point in the image and time instance in the spatial-temporal space. The node contains the visual features and position information so that the node can be treated as a 3D point with other modal information such as RGB information in point cloud data acquired by RGB-D camera.

In the graph, the nodes are connected to each other in a bi-directional manner. In summary, the above-mentioned graph is represented by Eq. (1).

$$\begin{gathered} G_e = (V_e, \mathcal{E}_e) \\ V_e = \{(\boldsymbol{f}_i, \boldsymbol{p}_i)\}_{i=1}^{|V_e|} \\ \boldsymbol{f}_i = RoIAlign(\boldsymbol{F}_i), \boldsymbol{p}_i = (x_i, y_i, t) \in R^3 \\ \mathcal{E}_e = \{e_{j,k}\} \end{gathered} \tag{1}$$

where G_e is a graph whose nodes and edges are V_e and $\mathcal{E}_e$, respectively. V_e has the visual features $\boldsymbol{f}_i$ and position vector $\boldsymbol{p}_i$, where $i(= 1, \ldots, |V_e|)$ is the index of a node. The visual feature is the output of $RoIAlign$ of feature map $\boldsymbol{F}_i$. The position vector $\boldsymbol{p}_i$ is the three-dimensional vector whose elements are the coordinates (x_i, y_i) of the centroid of the bounding box, and the time instance t in which the corresponding frame is observed. $e_{j,k}$ is the edge between j-th node and k-th node. As the time

instance t passes, the spatial-temporal graph grows as shown in Fig 4. Thus, the variation of the number and position of the point is represented by changes in the geometric shape of the graph spatial-temporal space.

3.3.2 Partial Spatial-Temporal Graph

Although the graph which can represent the shape in the spatial-temporal dimension is constructed, there is possibility that the geometric shape of the graph is not considered sufficiently because our model whose building block is Point Transformer Layers processes the entire structure of the graph and ignore the local feature. To deal with this issue, we divide the spatial-temporal graph into a few sub-graphs. Concretely, the sub-graphs are built in each equally-divided duration. In the following, the equally-divided duration is called "temporal segment". For example, in case of that the entire duration is 30 seconds with three temporal segments, the first, second and third temporal segments correspond to from 1 to 10 seconds, from 11 to 20 seconds and from 21 to 30 seconds respectively. The sub-graph is generated based on the frames in each temporal segment. If $t <$ 10 seconds, in the temporal segment corresponding to duration from 1 to 10, the spatial-temporal graphs is acquired by accumulating spatial-graphs over the temporal dimension: i.e. for accumulating the graphs over temporal dimension, the nodes obtained from the new observed frame are appended to the already built graphs, and the new nodes are connected to the nodes in the already built graph. Note that the spatial and temporal connections among the nodes is full and bi-directional.

In other temporal segments in which the current time instance is before the start of the temporal segment, the dummy graph, which contains only one node having a zero vector with self-connecting edge, is generated.

In summary, the above-mentioned processes are represented by Eq. (2).

$$
\begin{aligned}
G_{p,\Delta_l,\Delta_{l+1}} &= \begin{cases} (V_p, \mathcal{E}_p) & \text{if } \Delta_l \le t < \Delta_{l+1} \\ (V_z, \mathcal{E}_z) & \text{else } t < \Delta_l \end{cases} \\
V_p &= \{(\boldsymbol{f}_i, \boldsymbol{p}_i)\}_{i=1}^{|V_p|} \\
\boldsymbol{f}_i &= RoIAlign(\boldsymbol{F}_i), \boldsymbol{p}_i = (x_i, y_i, t) \in R^3 \\
\mathcal{E}_p &= \{e_{j,k}\} \\
V_z &= (\mathbf{0}_f, \mathbf{0}_p), \mathcal{E}_z = \{e_{1,1}\}
\end{aligned}
\tag{2}
$$

In Eq. (2), $G_{p,\Delta_l,\Delta_{l+1}}$ is a graph generated in the temporal segment which corresponds to temporal duration from Δ_l to Δ_{l+1}. If the current time instance

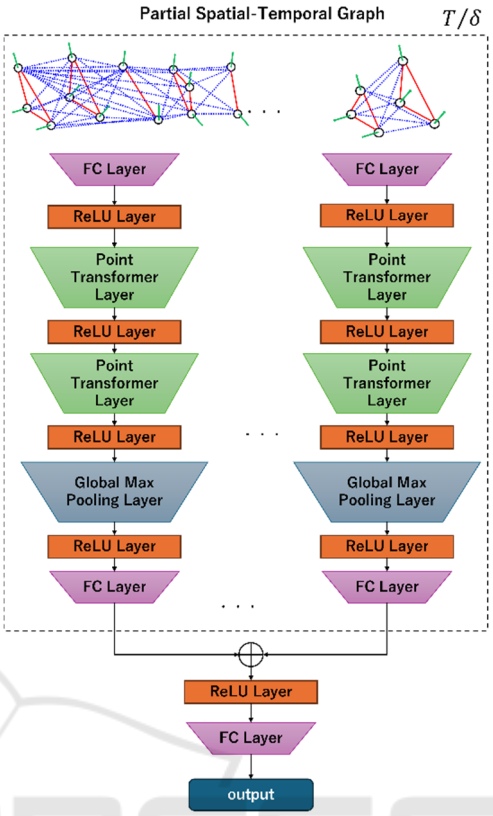


Figure 6: The structure of our model.

t is in the temporal segment, the graph is $(V_p, \mathcal{E}_p)$, while the graph $(V_z, \mathcal{E}_z)$ is the time instance t is before the start of Δ_l. Similar to Eq. (1), $(V_p, \mathcal{E}_p)$ is the graph whose node contains the visual feature of the bounding box surrounding the participant and the position vector including the centroid coordinate of the bounding box and the time instance. $V_z = (\mathbf{0}_f, \mathbf{0}_p)$ has the zero vectors whose dimensions are same as the visual feature and position vector, respectively. $\mathcal{E}_z = \{e_{1,1}\}$ means self-connection of only one node V_z. The process of generating partial spatial-temporal graph result in generating T/δ graphs where T is the length of entire duration and $\delta = \Delta_{l+1} - \Delta_l$ is the length of each temporal segment. By generating the spatial-temporal graph at each segment, the local geometric feature can be preserved.

3.4 Graph Processing

After the partial spatial-temporal graph are acquired, these graphs are input to our model. The input graphs have the two aspects; graph and set of points with visual features. Therefore, in this paper, the model should be built based on the basic blocks which can deal with these aspects.

The model in our proposed method is shown in Fig 6. The basic block is Point Transformer Layer from (Zhao et al., 2021) which takes the point cloud data as the input and deal with various task such as classification, semantic segmentation and part segmentation. Here, a brief review about point transformer layer is shown.

In Point Transformer (Zhao et al., 2021), the self-attention structure and position encoding are applied to point cloud data which is a set of 3D points. The processing in Point Transformer Layer is shown in Eq. (3)

$$\boldsymbol{x}'_i = \sum_{\boldsymbol{x}_j \in \chi(i)} \rho(\gamma(\varphi(\boldsymbol{x}_i) - \psi(\boldsymbol{x}_j) + \delta)) \odot (\alpha(\boldsymbol{x}_j) + \delta) \quad (3)$$

In Eq. (3), $\chi(i)$ is a set of points which are connected to $\boldsymbol{x}_i$ which is feature of i-th point. $\boldsymbol{x}'_i$ is the output of Point Transformer Layer of $\boldsymbol{x}_i$. ρ is a normalization function. γ is a non-linear function. φ, ψ and α are functions for point wise transformation. δ is position encoding function which is defined as below.

$$\delta = \theta(\boldsymbol{p}_i - \boldsymbol{p}_j) \quad (4)$$

In Eq. (4), θ is an encoding function. $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$ are the i-th and j-th position vectors, respectively.

As shown in Eq. (3) and (4), Point Transformer Layer processes both the feature and position of the point. The layer is used as building the block for classification and segmentation of the point cloud. Looking at our problem setting, our spatial-temporal graph, which expand as time goes, can be treated as the set of points with the visual features with the variation of the shape. The reason is that the number, position and actions of the participants vary over time. Therefore, we apply Point Transformer Layer to deal with our input graphs because the layer fits well to the form of our input graph. Note the implementation of the layer follows that of Pytorch Geometric (Pytorch Geometric, 2024).

Our model has multi stream network structure so that the local geometric structure of spatial-temporal graph can be utilized in classification. Each stream inputs the partial spatial-temporal graph from each temporal segment. The multi stream extracts the features from the graphs in the temporal segments respectively. The last layer of each stream is global max pool layer because the graph structure should be represented in the form of a vector. The extracted features from the streams are concatenated into one vector. Finally, the concatenated vector is fed into the small network for classification.

4 EXPERIMENTS

4.1 Dataset

To confirm the validity of our proposed method, we conducted experiments on the publicly released dataset called 4D-OR (Özsoy et al., 2022). The dataset contains videos capturing mock surgery of knee-replacement using multiple cameras. Using each camera, 10 videos were acquired. The number of participants is five. Note that our experiments do not deal with the patient, which (Özsoy et al., 2022)'s dataset includes; that is, our assumption is the patient is not the member of the group.

Additionally, our method and (Zhai et al., 2023) require bounding boxes surrounding the members of the group in all frames because both (Zhai et al., 2023) and our proposed method assume extracting the visual features from local areas which surround the participants in each frame. On the other hand, some of surgical processes in 4D-OR (Özsoy et al., 2022) have frames in which any of the participants does not appear or in which detecting the participants automatically is difficult even if the detection model is well-finetuned. Therefore, the surgical processes in which the participants always appear in all the frames and are easy to be detected in each frame with object detector such as (Liu et al., 2016) are chosen. For our experiment, the video acquired by the second camera of 4D-OR in the operating room is selected. The contents of the surgical processes are shown in Table 1, where, if necessary, refer to (Özsoy et al., 2022) or (Özsoy et al., 2024) for details of the dataset. The average duration of each surgical process is computed based on the number of the frames belonging to each surgical process.

Table 1: The content and the average duration of each surgical process.

Process	Average duration	# of instances
Knee- prep	01 min 56 sec	10
Knee-insert	01 min 42 sec	10
Patient-prep	02 min 55 sec	10
Surgery-conclusion	01 min 28 sec	10

4.2 Conditions

For building our model, seven out of the ten videos capturing surgeries are used for the training, one is used for the validation, and the other two are used for the test, where the one for validation is selected randomly from the eight patterns after the two test

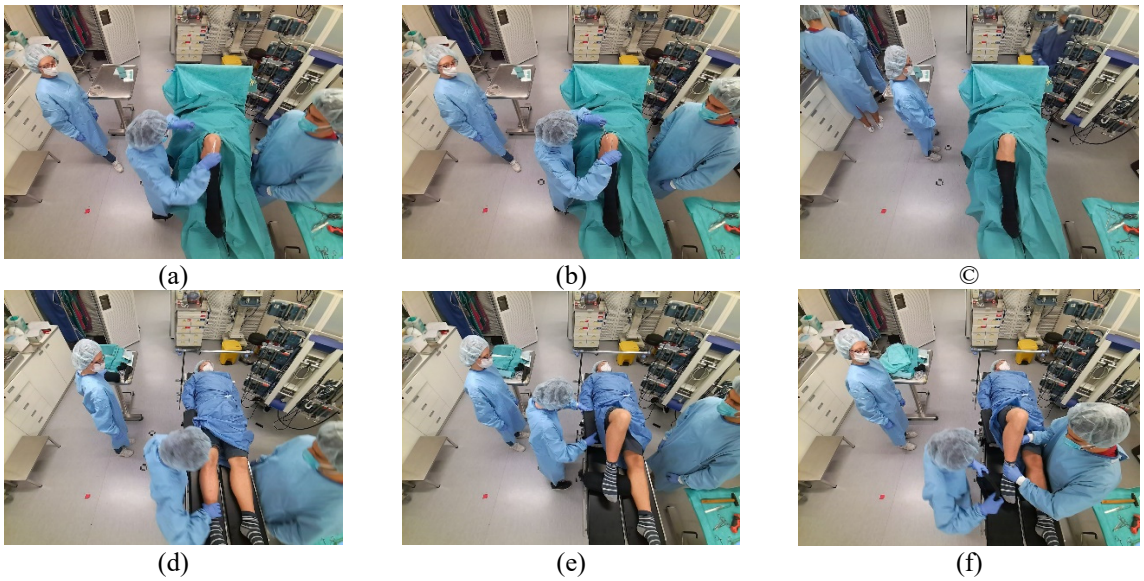


Figure 7: The examples of the images of [34] from surgical processes in which the passed durations are 1, 10 and 20 seconds. The upper row shows the images of the process called surgery conclusion. The lower rows shows the images of the process called patient-prep in which the passed durations are 1, 10 and 20 seconds.

patterns are removed. Therefore, the remaining seven are used for training. The 45 test patterns are examined in this paper.

For optimizing, Adam is used where the batch is 64.

In our experiments, the width and height of the image are scaled to 1/8 of the original size of images in (Özsoy et al., 2022). A machine used for conducting our experiments is Ubuntu 20.04, AMD® Ryzen 7 2700x eight-core processor x 16 and Nvidia Geforce RTX 3060 which has one 12GB video card.

In this experiment, VGG16(Simonyan et al., 2015) is used as the image feature extractor as shown in Fig. 3. The extractor is not finetuned in advance to train the entire model due to simplification of the experiments.

In our experiments, it is assumed that the participants are detected in images before both training and inference. The participants are detected by object detector SSD (Liu et al., 2016) which is finetuned with frames sampled from all over the dataset. For (Özsoy et al., 2022)'s dataset, in each frame of the video, whether the number of detected participants is equal to or less than the maximal number of the participants is checked. If the number of the detected participants is greater than the maximal number of the participants, the detections which do not comprise any of the participants are removed because both previous work (Zhai et al., 2023) and our proposed method assume extracting the visual features from the local area (bounding box) which comprises one participant in each frame. In case of no detection despite of the fact that (a) participant(s) is/are in the frame, the bounding boxe(s) is/are placed manually.

For training the model, 20 epochs for 4D-OR (Özsoy et al., 2022) are used. For training and inference, the early observation is set to the duration corresponding to from 1 to 30 seconds for all the surgical processes. Note that each participant's action is not treated in both training and inference.

4.3 Results

The results of the early recognition on 4D-OR (Özsoy et al., 2022) are shown in Table 2 respectively. Using the F-value of each process in each test, the means, and standard deviations of F-values all over the test patterns are computed.

Table 2: The f-value at each process of 4D-OR (Özsoy et al., 2022).

Process	Mean and standard deviation of F-value
Knee-prep	0.787 +/- 0.193
Knee- insert	0.852 +/- 0.104
Patient-prep	0.900 +/- 0.141
Surgery-conclusion	0.682 +/- 0.230

In this paper, the comparison with state of the art in terms of early recognition of the group activity is conducted to confirm the superiority of our proposed method. While there are two major conventional

works (Chen et al., 2020), and (Zhai et al., 2023) as stated in Sec 2, in this paper, (Zhai et al., 2023) is chosen for this comparison because (Zhai et al., 2023) is state of the art among the major works of group activity based early recognition and their codes are publicly released. For this comparison the method proposed by (Zhai et al., 2023) is implemented based on the available code and modified so that the dataset can be examined. Note that (Zhai et al., 2023) 's visual feature extractor is finetuned in advance to training the entire model.

The results shown in Table 3, in which F-value are listed. The bounding boxes surrounding the participants are detected by SSD (Liu et al., 2016), which is finetuned based on the datasets used in this paper respectively. For training and inference, the early observation is set to temporal lengths corresponding to from 1 to 30 seconds for all the surgical processes.

Table 3: The comparison with (Zhai et al., 2023) at each process of 4D-OR (Özsoy et al., 2022).

Process	(Zhai et al., 2023)	Ours
Knee- prep	0.653 +/- 0.260	0.787 +/- 0.193
Knee-insert	0.715 +/- 0.193	0.852 +/- 0.104
Patient-prep	0.907 +/- 0.133	0.900 +/- 0.141
Surgery-conclusion	0.636 +/- 0.234	0.682 +/- 0.230

5 DISCUSSIONS

5.1 Contribution and Limitations

As shown in Table 2, the F values is 68.2% at least based on the observed frames until 30 seconds from the beginning, which correspond to 34.1 % observation in surgery conclusion. Furthermore, the F values in knee-prep, knee-insert and patient-prep are more than 75% which is relatively high value. These results mean our model can recognize these surgical processes based on the video frames which are observed until 30 seconds. In terms of observation ratio, if 25.9% (almost equals to 30 secs / 1 min 56 secs), 29.4% (almost equals to 30 secs / 1 min 42 secs) and 17.1% (almost equals to 30 secs / 2 min 55 secs) observed information is obtained, our model can recognize these surgical processes. The important element of this successful performance seems to derive from representation of the input and the model design. In terms of the representation of the input, the spatial-temporal graph contains not only the visual features but also the position and the number of the participants and their temporal variation. Furthermore, in our input graph, the temporal variation of the position and the number of the participants is a geometric shape in spatial-temporal space. For example, if the number of participants changes from three to two, the number of vertices decreases. In the perspective of geometric shape, the shape of the point set is sharpened. The shape in spatial-temporal space can be distinguishable cue in early stage at high accuracy. In terms of model design, using Point Transformer Layer which can treat both the feature and position of the points fit to our input graph data.

On the other hand, the F-value at surgery-conclusion is lower than those of other processes. This comes from the similarity in terms of the shape of the graph in spatial-temporal space. Looking at the frames of surgery conclusion until 30 seconds, the position and number of participants seem to be similar to other processes for example the patient-prep until 20 seconds. As time passes, the unique cues come to appear in the video frames. The cues are for example that the surgeons get out of the room and remained assistants start to clear the instruments. However, our method seems to fail to capture this detailed cue. One of the possibilities of the failure lies in how to make the temporal segments for generating partial spatial-temporal graph. Although we took the strategy which divides the temporal duration to make multiple spatial-temporal graphs so that the shape feature of the graph can be preserved, it may be difficult to capture the local shape for our model if the granularity of the division of temporal segments is coarse. Therefore, how to divide the temporal segments for generating optimal graph from other processes should be researched as future work.

5.2 Comparison to the State of the Art

Table 3 shows that our proposed method performs better or competitive compared to (Zhai et al., 2023). In particular, our proposed method outperforms (Zhai et al., 2023) in the recognition of knee-prep and knee-insert significantly. The reason seems to be that our method treats not only the visual features of the participants but also the positions and the number of those by representing them as the shape of the graph, while (Zhai et al., 2023) treats only the visual features of the participants by merging them via pooling.

6 CONCLUSIONS

This paper has proposed a group activity-based method for early recognition of surgical processes using the camera attached to the ceiling of the

operating. To deal with this problem setting, the proposed method (1) makes the spatial-temporal graphs which represents not only visual features of the participants but also the positions and the number of the participants and (2) uses the graph to classify to the category of the surgical processes. Our model utilizes Point Transformer Layer as the building blocks to deal with the graphs which has the geometric information in spatial-temporal space.

By using the model, the early recognition of the surgical process is performed on the public datasets; mock surgery of knee replacement (Özsoy et al., 2022). Experimental results show that our method can recognize each surgical process from early durations of the inputted video, where the F1 values in the public dataset (Özsoy et al., 2022) are from 68.2% to 90.0% in 30 seconds from the beginning. These results mean our method can recognize the surgical process based on from early part 17.1 % to 34.1 % of the entire information in (Özsoy et al., 2022), respectively. Compared with the state-of-the-art in terms of early recognition for the group activity (Zhai et al., 2023), it is shown that that ours outperforms (Zhai et al., 2023) in 4D-OR dataset (Özsoy et al., 2022).

REFERENCES

Y. Li, J Ohya, T. Chiba, R. Xu, H. Yamashita (2016). ” Subaction Based Early Recognition of Surgeons' Hand Actions from Continuous Surgery Videos”, IIEEJ Transactions on Image Electronics and Visual Computing Vol.4 No.2

H Zhao, P. Wildes (2021). ” Review of Video Predictive Understanding: Early Action Recognition and Future Action Prediction”, arXiv:2107.05140v2.

B. Zhou, A. Andonian, A. Oliva, A. Torralba (2018). ” Temporal Relational Reasoning in Videos”, arXiv:1711.08496

L. Chen, J. Lu, Z. Song, J Zhou (2018). ” Part-Activated Deep Reinforcement Learning for Action Prediction”, Proc. of ECCV, pp 435–451.

G. Singh, S. Saha, M. Sapienza, P. Torr, F. Cuzzolin (2017). ” Online Real-Time Multiple Spatiotemporal Action Localisation and Prediction”, Proc of ICCV, pp. 3657–3666

C. Sun, A. Shrivastava, C. Vondrick, R. Sukthankar, K. Murphy, C. Schmid (2019). ” Relational Action Forecasting”, Proc of CVPR, pp. 273–283

S. Ma, L. Sigal, S. Sclaroff: (2016). ”Learning Activity Progression in LSTMs for Activity Detection and Early Detection”, Proc of CVPR, pp.1942–1950

M. S. Aliakbarian, F. S. Saleh, M. Salzmann, B. Fernando, L. Petersson, L. Andersson (2017). ” Encouraging LSTM to Anticipate Actions Very Early”, arXiv:1703.07023

Y. Kong, Z. Tao, Y. Fu (2017). ” Deep Sequential Context Networks for Action Prediction”, Proc of CVPR, pp.3662–3670

Y. Kong, D. Kit, Y. Fu (2014). ”A Discriminative Model with Multiple Temporal Scales for Action Prediction”, Proc of ECCV, pp. 596–611

X. Wang, J. Hu, J. Lai, J. Zhang, W. Zheng (2019). ”Progressive Teacher-Student Learning for Early Action Prediction”. Proc of CVPR, pp. 3551–3560.

J. Bradbury, S. Merity, C. Xiong, R. Socher (2016). “Quasi-Recurrent Neural Networks”, arXiv:1611.01576v2

H Zhao, R. Wildes (2019). ”Spatiotemporal Feature Residual Propagation for Action Prediction”, Proc of ICCV, pp. 7002–7011

K. Soomro, A. R. Zamir, M. Shah (2012). “UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild”, arXiv:1212.0402

H. Jhuang, J. Gall, S. Zuffi, C. Schmid, M. J. Black (2013). “Towards Understanding Action Recognition”, Proc of ICCV, pp. 3192–3199

Y. Kong, Y. Jia, Y. Fu (2012). ”Learning Human Interaction by Interactive Phrases”, Proc of ECCV, pp. 300–313,

A. Patron-Perez, M. Marszalek, A. Zisserman, I. Reid: “High Five: Recognising human interactions in TV shows”, Proc of BMVC, pp. 50.1–50.11, (2010) (doi:10.5244/C.24.50)

R. Goyal et al. (2017). “The “Something Something” Video Database for Learning and Evaluating Visual Common Sense”, Proc of ICCV, pp. 5843– 5851

J. Hu, W. Zheng, J. Lai, J. Zhang (2017). “Jointly Learning Heterogeneous Features for RGB-D Activity Recognition”. TPAMI, vol.39, no.11, pp.2186-2200

Y. Li, C. Lan, J. Xing, W. Zeng, C. Yuan, J. Liu (2016). ”Online Human Action Detection Using Joint Classification-Regression Recurrent Neural Networks”, proc of ECCV, pp. 203–220

C. Liu, Y. Hu, Y. Li, S. Song, J. Liu (2017). “PKU-MMD: A Large Scale Benchmark for Continuous Multi-Modal Human Action Understanding”, arXiv:1703.07475

K.Simonyan, A.Zisserman, (2015). “Very Deep Convolutional Networks for LargeScale Image Recognition”, arXiv:1409.1556v6,

J. Chen, W. Bao, Y. Kong: Group Activity Prediction with Sequential 14 Relational Anticipation Model, arXiv:2008.02441v1 (2020)

The Japan Institute for Labour Policy and Training (2022). https://www.jil.go.jp/kokunai/blt/backnumber/2022/11/s_02.html access 2023/04/

M Ibrahim, et al. (2016) “A Hierarchical Deep Temporal Model for Group Activity Recognition”, arXiv:1511.06040

W. Choi, K. Shahid, S. Savarese, (2009). “What are they doing? : Collective Activity Classification Using Spatio-Temporal Relationship Among People.” 9th International Workshop on Visual Surveillance (VSWS09) in conjuction with ICCV, https://cvgl.stanford.edu/projects/collective/collectiveActivity.html

W.Liu et al., (2016) “SSD: Single Shot MultiBox Detector”, arXiv:1512.02325

X. Zhai et al. (2023). "Learning Group Residual Representation for Group Activity Prediction*," 2023 IEEE International Conference on Multimedia and Expo (ICME), Brisbane, Australia, pp. 300-305,

K. Yokoyama et al. (2023). “Operating Room Surveillance Video Analysis for Group Activity Recognition,” Advanced Biomedical Engineering, vol.12, p.171-181,

E. Özsoy et al., “4D-OR: Semantic Scene Graphs for OR Domain Modeling”, arXiv:2203.11937

E. Özsoy et al., (2024). “Holistic OR domain modeling: a semantic scene graph approach”, Int J CARS 19, 791–799.

A. Shargi et al. (2020). “Automatic Operating Room Surgical Activity Recognition for Robot-Assisted Surgery”, arXiv:2006.16166

L. Bastian et al. (2022), “Know your sensORs -- A Modality Study For Surgical Action Classification”, arXiv:2203.08674

J. Wu et al. (2019). “Learning Actor Relation Graphs for Group Activity Recognition”, arXiv:1904.10117

S. Li et al. (2021). “GroupFormer: Group Activity Recognition with Clustered Spatial-Temporal Transformer”, arXiv:2108.12630

Microsoft (2021). https://learn.microsoft.com/ja-jp/previous-versions/azure/kinect-dk/hardware-specification#depth-camera-supported-operating-modes accessed at 2024/09/08

H. Zhao et. al, (2021) “Point Transformer”, arXiv:2012.09164

Pytorch Geometric (2024) PyG Documentation — pytorch_geometric documentation (access 2024/11/13)

K. He et al., (2017), Mask-RCNN, arXiv:1703.06870