

# Punch Type Classification and Hit Judgement Using Estimated Skeletal Model in Boxing Match Videos

Soma Watanabe<sup>1</sup> and Yoshinari Kameda<sup>2</sup><sup>a</sup>

<sup>1</sup>College of Engineering Systems, University of Tsukuba, Japan

<sup>2</sup>Center for Computational Sciences, University of Tsukuba, Japan

**Keywords:** Skeletal Description, Action Classification, Action Recognition, Boxing, Sports Video.

**Abstract:** In boxing, the choice of punch types and how to hit the punch to the opponent player is an important issue. So the support of computer vision on punch type classification and hit judgement on boxing match video is demanded. There are currently two challenges to that purpose. The first is the preparation of appropriate video dataset of boxing matches. The second is the discussion of the right method for punch type classification and punch hit judgment. We propose to create a video dataset of boxing matches from a boxing 3DCG simulation. The simulation can automatically annotate attributes to the dataset. This is useful for hit / no-hit judgement as it is not easy to identify which punches are actually hit and which are not on real boxing match videos. Based on the dataset we prepared, we propose a new method using time-series skeletal representation for classifying the type of punches and judging the hits. The experimental results show that our proposed method is able to classify the types of punches and judge the hits.

## 1 INTRODUCTION

Boxing is a combat sport in which victory or defeat is determined by two fighters exchanging punches. In boxing, the key to victory is to use different types of punches to hit the opponent. Automation of the classification of the type of punches and the judging of hits from boxing match videos would contribute to the development of tactical analysis in boxing.


There are two challenges in analyzing punches using computer vision. The first is to prepare a video data set of boxing matches with correct annotation. Due to the nature of boxing, the number of fights per fighter in boxing is not so large. In addition, in order to use the match videos as training data, it is necessary to assign not only the type of punch as a teacher signal but also the attribute of whether the punch was a hit or not. It is not easy to visually assign attributes to the video. To the best of our knowledge, there are no good boxing match video datasets that describe these attributes and are publicly available. Second, partly due to the first reason, no method has been proposed to automatically classify the type of punch and judge the hit based on boxing match video.

In this study, we propose a method to automatically recognize punches based on a boxing match video dataset created from a CG simulation of a boxing match. In automatic punch recognition, both type classification and hit judgment are performed.

In the generation of punch images in a boxing CG simulation, motion for each punch type is used. By performing the punch hit judgment on the simulation, the true value of the hit judgment attribute is automatically assigned to the dataset for each punch.

We propose an automatic recognition method using time-series skeletal representation for classifying the type of punch and judging the hit. The framework is the same for both classification and judgement. A learning model suitable for recognizing the time-series skeletal representation of multiple human bodies, including occlusions, is used for both. We formulate the punch type classification as a multi-class classification problem and the hit judgment as a two-class classification problem.

The contribution of this research is twofold. The first is the creation of a boxing match video dataset using CG simulation. We created a video dataset suitable for skeletal estimation by using 3DCG simulation of a scene of a boxing match in which a boxer is

 <https://orcid.org/0000-0001-6776-1267>

punching. The created dataset is used for both training and validation. Second, we show that the time-series skeletal representation of two fighters can be used to classify the type of punch and judge the hit.

The general flow of the proposed method is described below. First, a 3DCG animation is created. Next, punch motion is captured from various angles to create punch videos. A punch type label and a hit label are assigned to each punch video. In the training phase, we first apply skeletal estimation to the punch videos to obtain a skeletal time series representation for each punch video. Then, using this set of skeletal time-series representations, we train a model that classifies the type of punch. Similarly, we train a model that judges the punch hit. ST-GCN++(Duan et al. 2022) is used for the machine learning model.

## 2 RELATED WORKS

Cizmic et al. proposed a punch classification method using wearable sensors (Cizmic et al. 2023). In the field of image recognition, punch classification methods for shadow boxing (Kasiri et al. 2017) (Kasiri et al. 2015) have been proposed, while Broilvskiy et al. proposed an action recognition method on a video dataset of a single fighter shadow boxing (Broilvskiy et al. 2021). The above studies are aimed only at punch type determination and do not discuss punch hit determination.

So far, recognition methods using skeletal representations have been presented for single-person actions in daily activities (Xu et al. 2023) (Lee et al. 2023). Similarly, action recognition methods have been presented for the daily actions of multiple persons (Li et al. 2020) (Tu et al. 2021) (Chen et al. 2023). In these studies, the interaction between persons has been cited as a problem in multi-person action classification tasks. In the same presentation, a method with high robustness against interactions was also reported.

Action classification methods using skeletal estimation have been reported for stand-alone sports such as skating (Li et al. 2021), taekwondo (Luo et al. 2023), and karate (Guo et al. 2021). These studies have shown that action classification methods using skeletal representations can provide superior recognition performance.

CG has been used as a dataset for supervised learning in the past, and Wood et al. have trained on CG-generated face image data alone for tasks such as identifying landmarks and analyzing regions of the human face, showing that training on CG face image data alone can record the same accuracy as training on actual face image data (Wood et al. 2023). They

have shown that when training with only CG face image data, the accuracy is comparable to that of training with actual face image data.

## 3 BOXING MATCH VIDEOS

In this study, boxing match video refers to video footage of two fighters punching each other with the goal of hitting a punch. It is assumed that the video may be shot from various positions around the ring. The camera work is assumed to be stationary when filming from each location, so as not to lose the generality of the filming.

The punch type judgment task in this study focuses on four types of punches used in boxing: hooks, jabs, straights, and uppercuts. In the punch hit judgment task, two types of punches are considered: hit and miss.

A single punch video is defined as a video of one punch by one of the two players in the match video.

For each punch video, we judge whether it is one of the four types of punches. For each punching video, and whether it is a "hit" or a "miss". In this study, we do not classify which player punched which.

A CG simulation is used to generate a number of punching videos. In one punching video, one of the two players performs the motion of punching and the other performs the motion of receiving a punch. For each punch video, two labels are assigned: punch type and punch hit. The punch type label is determined from the motion used. The label for the hit is automatically set based on the contents of the CG simulation.

## 4 DETERMINATION OF PUNCH TYPE AND HIT DETECTION

### 4.1 Video Processing

The punching video is acquired by CG simulation at 30 fps. For each frame in the punching video, we estimate the skeleton of two boxers using the 25 keypoint model of OpenPose (Cao et al. 2017). A time-series skeletal representation is obtained for the 17 keypoints (Figure 1) necessary for boxing action recognition. This time-series skeletal representation is used as input to the learning network model for judging the type of punches and hits. Figure 2 shows the processing steps. ST-GCN++(Duan et al. 2022) is used for both learning network models.

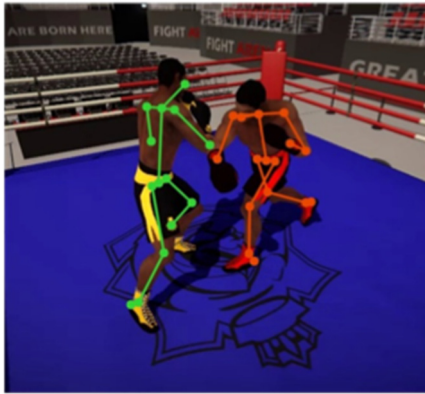


Figure 1: Estimated two skeleton models of two boxers. One skeleton model is composed of 17 key points.

## 4.2 Model

The punch classification task and the punch hit decision task in this study use ST-GCN++(Duan et al. 2022) as the learning network model, which is an effective model for skeleton-based action recognition tasks using graph convolution. ST-GCN++ is robust to person-person interactions. ST-GCN++ is also robust to occlusion caused by hiding of person regions by objects or persons. The model achieved an accuracy of 0.83 on the skeleton-based action recognition task on the NTURGB+D 120 dataset (Liu et al. 2019), which is a video dataset of basic everyday human actions.

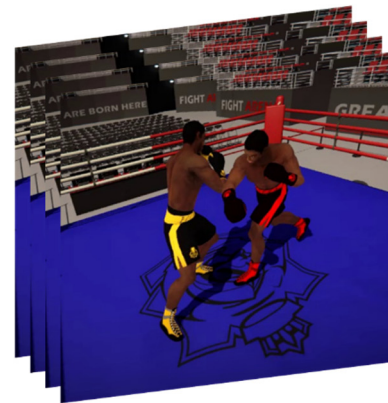
We use a trained model with the NTURGB+D 120 skeletal time series information dataset (Liu et al. 2019) from ST-GCN++. Fine tuning of the trained model is performed using the dataset created in this study.

## 5 VIDEO DATASET GENERATION

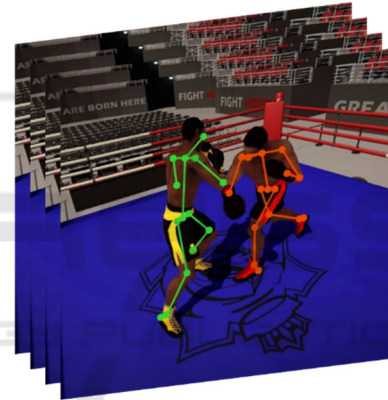
### 5.1 Camera Layout

Place the camera on one of the points on the sphere which is aligned to the center of the ring. The camera is positioned every 15 degrees horizontally and every 15 degrees vertically. The total number of shooting points is 96. The camera is set so that it faces the center of the ring at any point. Figure 3 shows a schematic of the camera arrangement.

To obtain a single punch image, one of the 96 locations is selected at random. To ensure the reproducibility of the experiment, the random number sequence is fixed.



Skeletal models estimation



ST-GCN++

ST-GCN++

Punch type label

Punch Hit (T/F)

Figure 2: Process flow of proposed punch type classification and hit judgement.

### 5.2 Punch Video Synthesis

Figure 4 shows the setup of the stadium model and boxer model used in this study. Two players are assigned to an offensive motion and a defensive motion for each. The offensive player randomly selects one of nine different punching motions, and the defensive player randomly selects one of 12 different defensive motions. All the motions were obtained as a commerc-

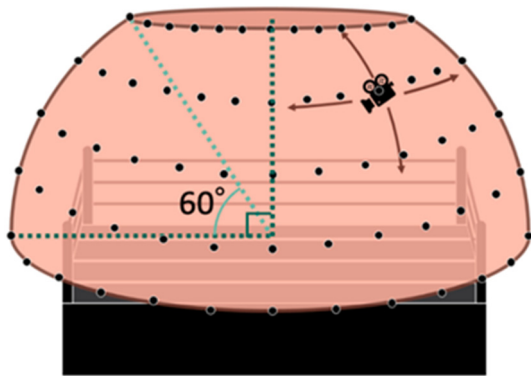


Figure 3: Layout of 96 Camera locations placed around the boxing ring.



Figure 4: A view of one camera location, showing the boxer CG models and the stadium model.

ial product of the motion dataset performed by experienced boxing athletes. The punch motions consist of three jabs, two each of hooks, uppercuts, and straights. Offensive and defensive selection, punch motion selection, and defensive motion selection are conducted randomly. During the experiment, the seed of the random number generation is fixed.

Since there are 9 punching motions, 12 defensive motions, and 96 camera positions, there are 10,368 punching videos. Using a random number sequence, 6,000 punch videos are used for training and 900 for validation. Table 1 and Table 2 show the distribution of punches of type classification and hit judgment in the training dataset created in this study. The average number of frames indicates the length of the motion as the video is recorded at 30fps. Figure 5 shows a part of the created punch videos. Figure 6 shows the results of the skeletal representation.

### 5.3 Hit Label Annotation

Collider in our 3D simulation is used to add a label to the punch hit decision. In this research, a sphere-shaped Sphere Collider is used for the glove and a capsule-shaped Capsule Collider is used for the player's body. The installation is shown in Figure 7. When the Sphere Collider on the hand of an offensive player



Figure 5: Four examples of generated punch motions.

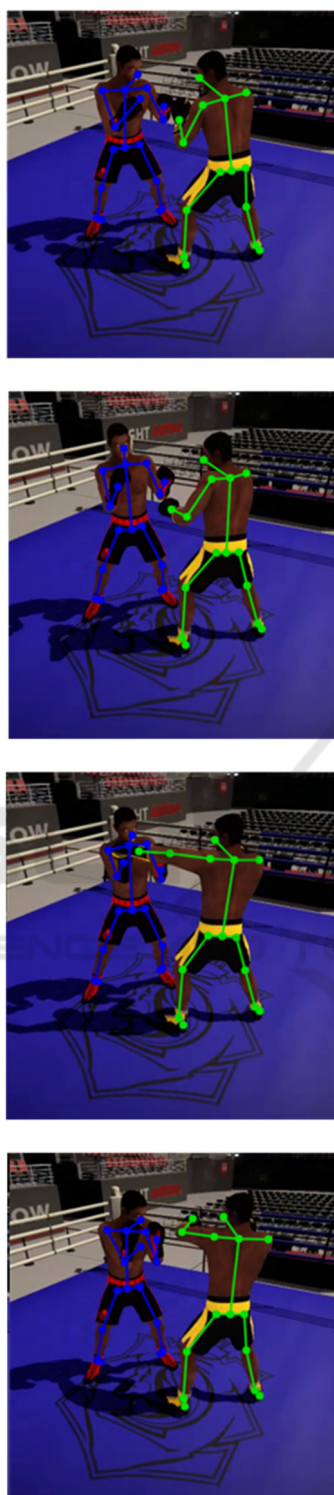


Figure 6: Skeletal model description of the punch motions of Figure 5.

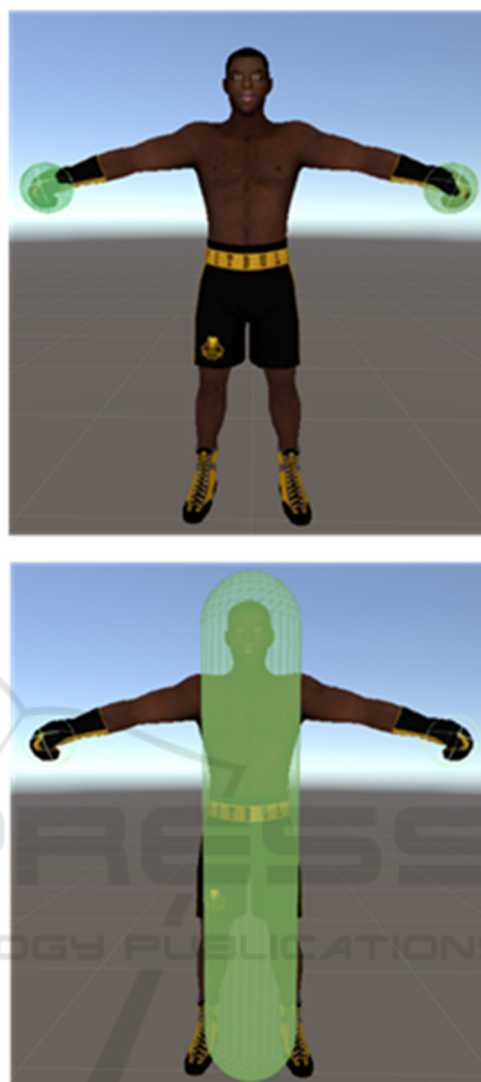


Figure 7: Layout of sphere collider at hands and capsule collider at body. Punch hit is detected by the collision of these colliders.

Table 1: Distribution of punch motions by types.

	hook	jab	staight	uppercut
#data	1376	2046	1273	1305
ratio	0.229	0.341	0.212	0.218
Average # frame	40.91	44.34	41.18	49.34

Table 2: Distribution of punch motions by hit / no-hit.

	hit	No hit
#data	2389	3611
ratio	0.398	0.602
Average # frame	43.61	44.41

collides with the Capsule Collider on the body of a defender, the label "hit" is added. When the hit does not occur, the label "no hit" is assigned.

## 6 EXPERIMENTS

For the punch type determination, 1,200 randomly generated punch videos were used for verification, yielding an accuracy of 0.94. Figure 8 shows the confusion matrix. The accuracy for each label exceeded 0.90, and a high accuracy of 0.98 was recorded for straight punches.

For the punch hit evaluation, 1,200 randomly generated punch videos were also used for verification, yielding an accuracy of 0.88. Figure 9 shows the confusion matrix. The accuracy for each label was 0.87 or higher.

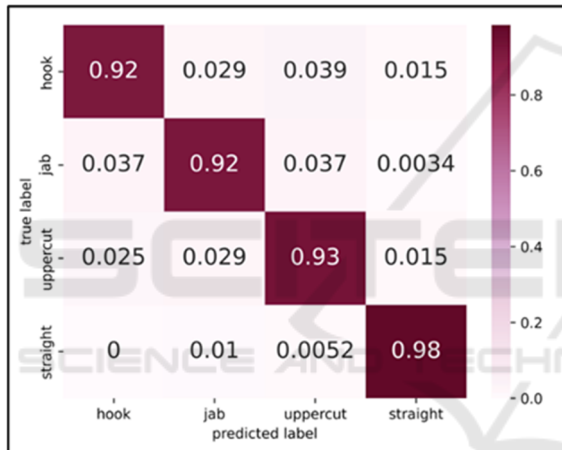


Figure 8: Confusion matrix of punch type classification.

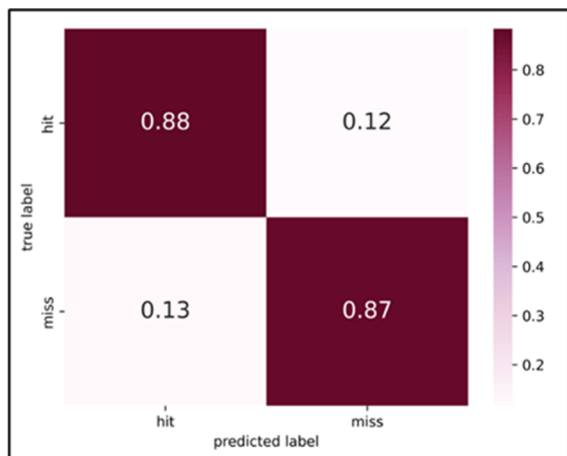


Figure 9: Confusion matrix of punch hit judgement.

## 7 CONCLUSIONS

In this paper, a boxing match video dataset was created from a boxing CG simulation, and an automatic punch recognition method was proposed based on this dataset. The automatic punch recognition is decomposed into two tasks: a task for classifying the type of punch and a task for judging the hit. For each of these tasks, we trained on ST-GCN++. Experimental results showed that the recognition rate for the punch type classification task was 0.94, and the recognition rate for the punch hit judgment task was 0.88.

This work was partially supported by JSPS KAKENHI 22K19803 and it was originated by (Watanabe, 2024).

## REFERENCES

- Duan, H., Wang, J., Chen, K., & Lin, D. (2022). Towards good practices for skeleton action recognition", Proc. the 30th ACM International Conference on Multimedia, 7351-7354.
- Cizmic, D., Hoelbling, D., Baranyi, R., Breiteneder, R., & Grechenig, T. (2023). "Smart boxing glove "RD  $\alpha$ ": IMU combined with force sensor for highly accurate technique and target recognition using machine learning," J. Applied Sciences, 13(16), 9073-9088.
- Kasiri, S., Fookes, C., Sridharan, S., & Morgan, S. (2017). "Fine-grained action recognition of boxing punches from depth imagery," J. Computer Vision and Image Understanding, 159, 143-153.
- Kasiri, S., Fookes, C., Sridharan, S., Morgan, S., & Martin, T. (2015). "Combat sports analytics: Boxing punch classification using overhead depth imagery," Proc. IEEE International Conference on Image Processing (ICIP), 4545-4549.
- Broilvskiy, A., & Makarov, I. (2021). "Human action recognition for boxing training simulator," Proc. 9th Analysis of Images, Social Networks and Texts (AIST), LNCS 12602, 331-343.
- Xu, H., Gao, Y., Hui, Z., Li, J., & Gao, X. (2023). "Language knowledge-assisted representation learning for skeleton-based action recognition," <https://arxiv.org/abs/2305.12398>.
- Lee, J., Lee, M., Lee, D., & Lee, S. (2023). "Hierarchically decomposed graph convolutional networks for skeleton-based action recognition," Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 10410-10419.
- Li, T., Sun, Z., & Chen, X. (2020). "Group-skeleton-based human action recognition in complex events," Proc. the 28th ACM International Conference on Multimedia, 4703-4707.
- Tu, H., Xu, R., Chi, R., & Peng, Y. (2021). "Multiperson interactive activity recognition based on interaction relation model," Hindawi Journal of Mathematics, 1-12.

- Chen, Z., Wang, H., & Gui, J. (2023). "Occluded skeleton-based human action recognition with dual inhibition training," Proc. the 31st ACM International Conference on Multimedia, 2625–2634.
- Li, H., Lei, Q., Zhang, H., Du, J., & Gao, S. (2021). "Skeleton-based deep pose feature learning for action quality assessment on figure skating videos," Proc. the 11th International Conference on Information Technology in Medicine and Education (ITME), 196-200.
- Luo, C., Kim, S., Park, H., Lim, K., & Jung, H. (2023). "Viewpoint-agnostic taekwondo action recognition using synthesized two-dimensional skeletal datasets," J. Sensors, 23(19), 8049-8063, 2023.
- Guo, J., Liu, H., Li, X., Xu, D., & Zhang, Y. (2021). "An attention enhanced spatial-temporal graph convolutional LSTM network for action recognition in karate," J. Applied Sciences, 11(18), 8641-8653.
- Wood, E., Baltrusaitis, T., Hewitt, C., Dziadzio, S., Cashman, T. J., & Shotton, J. (2023). "Fake it till you make it: Face analysis in the wild using synthetic data alone," Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 3681-3691.
- Cao, Z., Simon, T., Wei, E. S., Sheikh, Y. (2017). "Realtime multi-person 2D pose estimation using part affinity fields," Proc. IEEE conference on Computer Vision and Pattern Recognition (CVPR), 1302-1310.
- Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L., Kot, A. (2019). "NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding", Trans. IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI), 42(10), 2684-2701.
- Watanabe, S., Kameda, Y. (2024). "Automatic punch classification using skeletal estimation in boxing match videos," IEICE Tech. Rep., vol. 123, no. 433, MVE2023-83, pp. 224-228.