Human Action Recognition without Human

Yun He, Soma Shirakabe, Yutaka Satoh, Hirokatsu Kataoka
National Institute of Advanced Industrial Science and Technology (AIST)
Tsukuba, Ibaraki, Japan
{yun.he, shirakabe-s, yu.satou, hirokatsu.kataoka}@aist.go.jp

Abstract

The objective of this paper is to evaluate “human action recognition without human”. Motion representation is frequently discussed in human action recognition. We have examined several sophisticated options, such as dense trajectories (DT) and the two-stream convolutional neural network (CNN). However, some features from the background could be too strong, as shown in some recent studies on human action recognition. Therefore, we considered whether a background sequence alone can classify human actions in current large-scale action datasets (e.g., UCF101).

In this paper, we propose a novel concept for human action analysis that is named “human action recognition without human”. An experiment clearly shows the effect of a background sequence for understanding an action label.

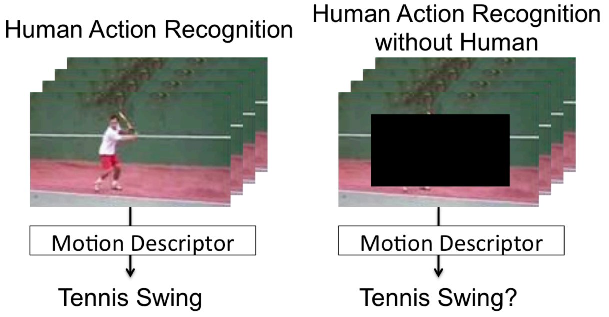


Figure 1: Human action recognition (left) and human action recognition without human (right): We simply replace the center-around area with a black background in an image sequence. We evaluate the performance rate with only the limited background sequence as a contextual cue.

1 Introduction

An effective motion representation is in demand for action recognition, event recognition, and video understanding. In human action recognition especially, several survey papers have been published in the last two decades [1, 2, 10, 11]. We have investigated a more reliable and faster algorithm to put action recognition into practice. The target applications of action recognition can be easily imagined, for example, surveillance, robotics, augmented reality, and intelligent surgery. However, current vision-based video representations focus on the media to improve the recognition rate on UCF101 [16], HMDB51 [7], and ActivityNet [4].

Here we categorize action recognition into two types: direct and contextual approaches.

The direct approach, which is motion representation, has been studied in action recognition. Since Laptev *et al.* proposed space-time interest points (STIP) [9, 8], xyt keypoint acquisition has been well established in temporal representation. STIP is significantly improved with densely connected keypoints in the dense trajectories approach (DT) [18, 17]. The DT is a more natural approach for understanding whole body motions because it uses a large amount of tracked keypoints. Recently, two-stream CNN has been applied as a representative method in action recognition [15]. The two-stream convolutional neural network (CNN) uses spatial and temporal streams to extract appearance and motion features from RGB and optical flow input. The classification scores at each stream are fused for evaluating an objective video. Other CNN-based approaches apply a dynamic scene descriptor such as a pooled time series (PoT) [14] and capture sensitive motion with a subtle motion descriptor (SMD) [6].

The contextual approach is focused around the region of a human and can provide an important cue to improve human action recognition. In related work, Jain *et al.* [5] and Zhou *et al.* [20] showed that object and scene context aid in the recognition of human actions. Jain *et al.* carried out an evaluation of how much object usage is needed for action recognition [5]. They combined object information with a classifier score into the improved DT (IDT) plus Fisher vectors (FVs) [12] as a motion feature from a human area. A large number of object labels (15,923 objects), e.g., computer and violin, are corresponded to an output function with AlexNet as an object prior. The response of CNN-based object information must be combined with a motion vector for a richer understanding of human actions. In their evaluation, motion + object vector allow us to obtain a better feature in an image sequence. When using object

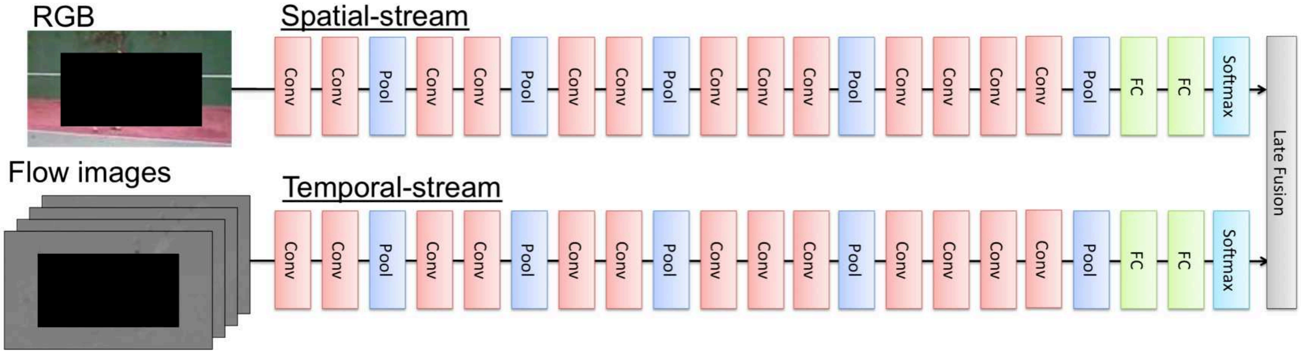


Figure 2: Very deep two-stream CNN [19] for human action recognition without human.

information, the performance rate rises by +3.9%, +9.9%, and 0.5% on UCF101, the THUMOS14 validation set, and KTH, respectively. According to experiments, the object vector improves recognition accuracy on a large-scale action database. Zhou *et al.* proposed a combination of a contextual human-object interaction and a motion feature for fine-grained action recognition [20]. Object proposals are captured by using BING [3]. However, some useless proposals are generated around a human area. The pruning of extra regions is executed by referring to dense trajectories around object proposals. The recognition rate can be improved with human-object interaction as a mid-level feature. The mid-level feature records an outstanding rate 72.4% on the MPII cooking dataset [13], which is known as a fine-grained action database. These two examples are convincing enough to integrate a mid-level feature into a motion vector. The mid-level feature including objects and backgrounds are enough to describe the situation around human(s).

The conventional approaches have implemented video-based human action recognition from a whole image sequence including a background. However, a curious option appears:

- Human action recognition can be done just by analyzing motion of the background.

To confirm this option, we try to prove the importance of the background on a well-studied dataset [16].

In this paper, we evaluate the effect of the background in human action recognition (see Figure 1). Our target is to measure a video-based recognition rate with a separated human and background sequence. We employ two-stream CNN [15] as a motion descriptor, and center-around image filtering to blind the human area.

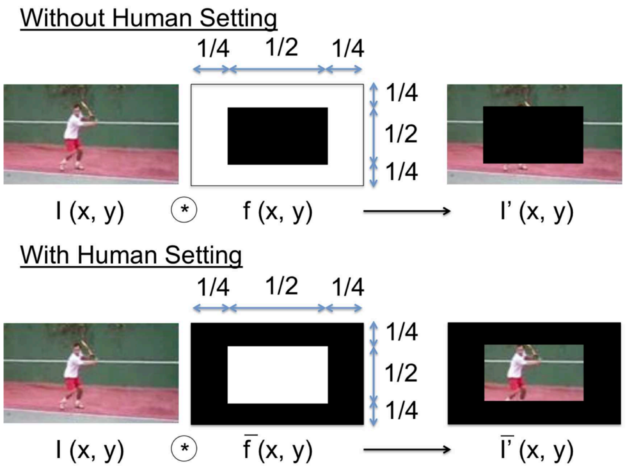


Figure 3: Image filtering for human action recognition without human.

2 Human Action Recognition *without* Human

The flowchart of human action recognition without human is shown in Figure 2. The recognition framework is based on the very deep two-stream CNN [19]. We only look at the appearance and motion features of the background sequence.

Setting without a human (see Figure 3 top). In the setting without a human, we calculate the image filtering with a black background as follows:

$$I^{'}(x,y) = I(x,y)*f(x,y) \qquad (1)$$

where $I^{'}$ and I show the filtered and input images, respectively, and x, y are pixel elements. Filter f replaces the

Stream	% on UCF101 (split 1)
Spatial stream	74.86
Temporal stream	80.33
Two-stream (S+T) [19]	84.30

Table 1: Performance rate on the UCF101 dataset with baseline two-stream CNN

center-around area with a black background. (The black background is a controversial representation.) The detailed operation is shown at the top of Figure 3.

Setting with a human (see Figure 3 bottom). We confirm the importance of the human appearance and motion features from an image sequence as follows:

$$\overline{I'}(x,y) \quad = \quad I(x,y) * \overline{f}(x,y) \qquad (2)$$

where $\overline{I'}$ and filter $\overline{f}$ are an inverse image and filter in the setting without a human. The background is eliminated with the inverse filter at the bottom of Figure 3.

Training of two-stream CNN. The learning parameters of the spatial and temporal streams are based on [19]. Our goal is to predict the video label without additional training in the setting without a human (see Figure 1). By using an original pre-trained model [19], we obtained the following results on UCF101 split 1: 74.86% (spatial), 80.33% (temporal), and 84.30% (two-stream) [1], as shown in Table 1.

3 Experiment

Dataset. We apply the well-studied UCF101 dataset. This large-scale dataset was mainly collected from YouTube videos of sports and musical instrument performance scenes. The recognition task is to predict an action label from a given video. The dataset contains several computer vision difficulties, e.g., camera motion, scaling, posture change, and viewpoint difference. The mean average accuracy is calculated with three training and test splits. Here we calculate an average precision with training/test split 1.

Quantitative evaluation. Table 2 shows the performance rate on the UCF101 dataset with or without a human. Surprisingly, the two-stream CNN performance was 47.42% in the setting without a human. We understand that a motion recognition approach relies on a background sequence. The spatial stream is +18.53% better than the temporal stream. Therefore, an appearance tends to classify between backgrounds. Motion features contribute slightly to the background classification; that is, the performance rate is increased +2.09% with the temporal stream. The two-stream CNN recorded 56.91% in the with human setting, which is +9.49% higher than the setting without a human.

Qualitative dataset evaluation. Figure 4 shows examples without a human setting on the UCF101 dataset. Where we evaluated partial and complete images without a human (Figures 4(a) and 4(b), respectively), the number of partial images without a human was 1,114 in 3,783 videos. The rate was 29.45% in UCF101 split 1. The complete images without a human were not found on the videos.

4 Conclusion

To the best of our knowledge, this is the first study of human action recognition without human. However, we should not have done that kind of thing. The motion representation from a background sequence is effective to classify videos in a human action database. We demonstrated human action recognition in with and without a human settings on the UCF101 dataset. The results show the setting without a human (47.42%) was close to the setting with a human (56.91 %). We must accept this reality to realize better motion representation.

References

[1] J. K. Aggarwal and Q. Cai. Human motion analysis: A review. Computer Vision and Image Understanding (CVIU), 1999.

[2] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. ACM Computing Survey, 2011.

[3] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr. Bing: Binarized normed gradients for objectness estimation at 300fps. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[4] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. Activitynet: A large-scale video benchmark for human activity understanding. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

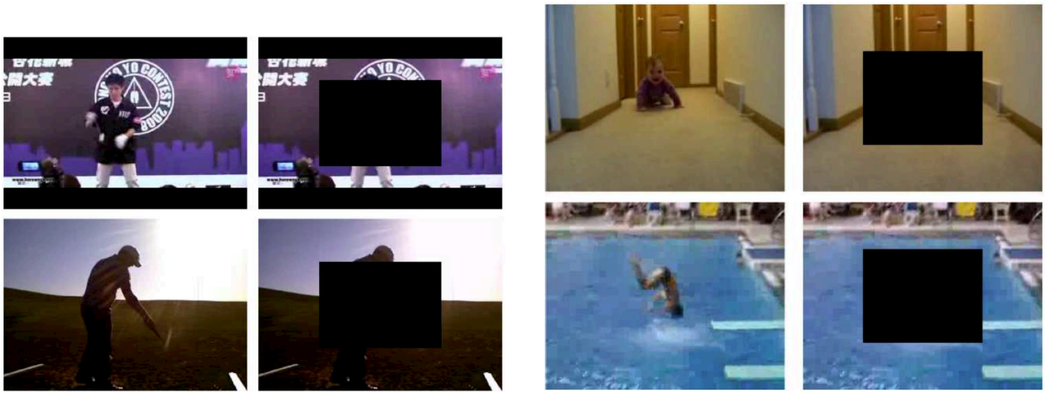
[5] M. Jain, J. C. van Gemert, and C. G. M. Snoek. What do 15,000 object categories tell us about classifying and localizing actions? IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[6] H. Kataoka, Y. Miyashita, M. Hayashi, K. Iwata, and Y. Satoh. Recognition of transitional action for short-term action prediction using discriminative temporal cnn feature. British Machine Vision Conference (BMVC), 2016.

[1] Our implementation is different from the report of Wang [19]. The performance rate depends on the parameter tuning. They reported 79.8% (spatial), 85.7% (temporal) and 90.9 % (two-stream) on UCF101 split 1.

With or Without a Human	Stream	% on UCF101 (split 1)
With human	Spatial stream	51.26
	Temporal stream	40.50
	Two-stream	**56.91**
Without human	Spatial stream	45.33
	Temporal stream	26.80
	Two-stream	**47.42**

Table 2: Performance rate of human action recognition with or without a human



(a) Partial without a human

(b) Entire without a human

Figure 4: Qualitative evaluation of the setting without a human on the UCF101 dataset.

[7] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. International Conference on Computer Vision (ICCV), 2011.

[8] I. Laptev. On space-time interest points. International Journal of Computer Vision (IJCV), 2005.

[9] I. Laptev and T. Lindeberg. Space-time interest points. International Conference of Computer Vision (ICCV), 2003.

[10] T. B. Moeslund, A. Hilton, and V. Kruger. A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding (CVIU), 2006.

[11] T. B. Moeslund, A. Hilton, V. Kruger, and Sigal L. Visual analysis of humans: Looking at people. Springer, 2011.

[12] F. Perronnin, J. Sanchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. European Conference on Computer Vision (ECCV), 2010.

[13] M. Rohrbach, S. Amin, Andriluka M., and B. Schiele. A database for fine grained activity detection of cooking activities. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[14] M. S. Ryoo, B. Rothrock, and L. Matthies. Pooled motion features for first-person videos. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[15] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition. Neural Information Processing Systems (NIPS), 2014.

[16] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human action classes from videos in the wild. CRCV-TR-12-01, 2012.

[17] H. Wang, A. Klaser, and C. Schmid. Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision (IJCV), 2013.

[18] H. Wang, A. Klaser, C. Schmid, and L. Cheng-Lin. Action recognition by dense trajectories. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011.

[19] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao. Towards good practices for very deep two-stream convnets. arXiv pre-print 1507.02159, 2015.

[20] Y. Zhou, B. Ni, R. Hong, M. Wang, and Q. Tian. Interaction part mining: A mid-level approach for fine-grained action recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.