Deep Key Frame Extraction for Sport Training

Meng Jian[1], Shijie Zhang[1], Xiangdong Wang[2(✉)], Yudi He[1], and Lifang Wu[1]

[1] School of Information and Communication Engineering,
Beijing University of Technology, Beijing 100124, China
[2] Sports Science Research Institute of the State Sports General Administration,
Beijing 10000, China
908583913@qq.com

Abstract. For some professional sports, it is required to supervise and analyze the athletics pose in training of athletes. In order to facilitate the browse of training videos, it's necessary to extract the key frames from training videos. In this paper, we propose a deep key frame extraction method for analyzing sport training videos. To alleviate the bias from complex background, Fully Convolutional Networks (FCN) is employed firstly to extract the foreground region which contains the athlete and barbell. Then over the extracted foregrounds, Convolutional Neural Networks (CNN) are leveraged to estimate the pose probability of each frame and extract the key frames by the maximum probability on each pose. The experimental results demonstrate that the proposed method achieves good performance in key frame extraction of sport videos comparing method.

Keywords: Key frame extraction · Pose estimation · Sport video
Fully Convolutional Networks (FCN)
Convolutional Neural Networks (CNN)

1 Introduction

With the development of sports, effective analysis of sport training becomes increasingly important. In some sports, it's necessary to analyze athletes' pose accurately in continuous actions. For example, in weight lifting, there are four key poses: (1) the athlete picks up the barbell to knees in pose of squat. (2) the athlete lifts the barbell and extend the knees. (3) the barbell is lifted up rapidly. (4) the athletes squat and lift the barbell to the top. Figure 1 provides examples of the key poses. Just like weight lifting, every sport has the unique key poses. Moreover, even in the same sport, different athlete acts different to some degree

X. Wang—This work was supported in part by Beijing Municipal Education Commission Science and Technology Innovation KZ201610005012, in part by the China Postdoctoral Science Foundation funded project 2017M610026 and in part by Project 1602 supported by the Fundamental Research Founds for the China Institute of Sport Science.

J. Yang et al. (Eds.): CCCV 2017, Part III, CCIS 773, pp. 607–616, 2017.
https://doi.org/10.1007/978-981-10-7305-2_51

on the same pose. Therefore, analyzing the video scientifically is a promising way to improve training quality [9].

In this work, we aim to extract the key frames from the training videos for the coach who is professional in supervising athletes in sports. Automatic key frame extraction would facilitate the coaches work and assist athlete improvement in professional actions. Therefore the extracted key frames should contain the key pose of athlete in sport videos. However, the contents in sports videos are full of complex background with a big scope of variations.

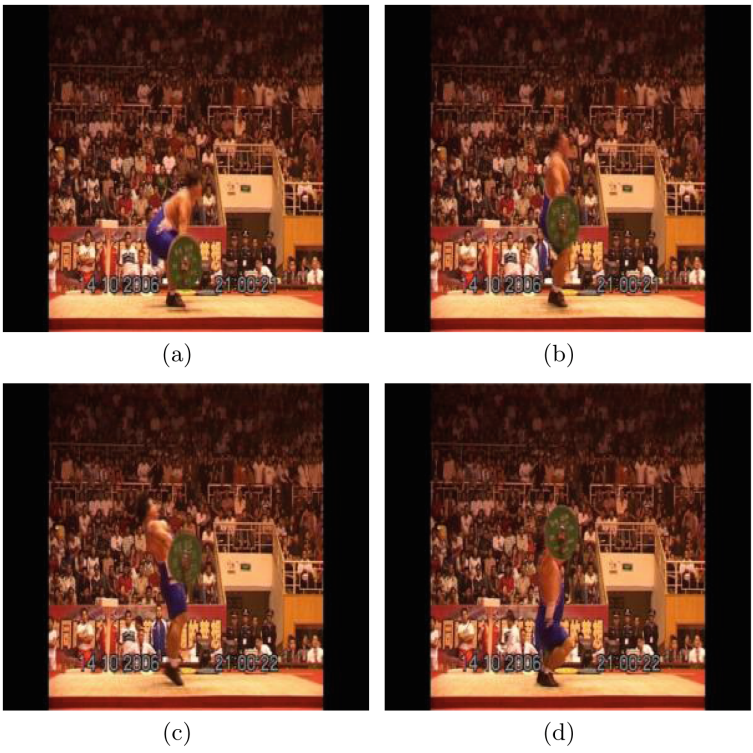


Fig. 1. Examples of key poses. (a) The athlete picks up the barbell to knees in pose of squat. (b) The athlete lifts the barbell and extend the knees. (c) The barbell is lifted up rapidly. (d) The athletes squat and lift the barbell to the top.

In some cases, the background of much audience becomes a big obstacle to focus on the characteristics of athlete in the video. In addition, there may exist some more barbells lying on the floor which also tends to make confusion to extract features of the barbell with the athlete. Figure 2(a) provides an example of complex background with many audiences and advertisement boards. Moreover, the other moving objects with various shifts also make confusion to analyze the key pose of the current frame. As an example, Fig. 2(b) illustrates a moving

(a) (b)

Fig. 2. Representative examples of complex background in sport competition videos. (a) There are many audiences in the auditorium and advertisement boards in front of the auditorium. (b) The air fan is running, and the shift of the air fan between frames is larger than that of the athlete.

fan that takes larger shifts than the key athlete. It implies that optical flow based techniques [10] are not proper to capture the corresponding characteristics of the key athlete for pose analysis and key frame extraction. How to deal with the bias from complex background is challenging for key frame extraction methods. The neighboring frames with similar contents also make obstacle to extract the key frames for different poses as examples in Fig. 3.

(a) (b) (c) (d)

Fig. 3. Examples of neighboring frames with similar contents, especially similar key pose of athlete.

Some researchers employed motion information in analyzing actions of the videos [6–8]. Laptev et al. [6] present a method for video classification that builds upon and extends several recent ideas including local space-time features, space-time pyramids and multi-channel non-linear SVMs. The algorithm [7] is able

to recognize and localize multiple actions in long and complex video sequences containing multiple motions. Schuldt et al. [8] demonstrate how such features can be used for recognizing complex motion patterns. However, the other moving objects also bring motion information. The motion information is not enough to distinct athlete and help estimate his poses. As a special case, key frame extraction of sport training videos is transferred to pose estimation and key pose extraction of frames. Recently, a variety of deep learning based methods [1–3] have been developed for human pose estimation. Ng et al. [1] fused raw frames and optical flow features to estimate poses of videos. Fan et al. [2] integrated both the local (body) part appearance and the holistic view of each local part for human pose estimation. Cheron et al. proposed pose-based convolutional neural network descriptor (P-CNN) [3] using different human body parts for action recognition. Inspired by these works, we employ deep neural networks of FCN and CNN to estimate poses of sport training videos, which aims to guide the training procedure and assist the coach to correct poses of athlete in professional sports.

In this work, we propose deep key frame extraction (DKFE) method for sport training videos. DKFE first extracts the foreground of athlete and barbell by FCN to alleviate the bias from complex background. Then we estimate key pose probabilities of each frame to extract the frames of key poses with the maximum probability. The main contributions of the proposed deep key frame extraction method are summarized as follows:

- Pose analysis of frames is constructed over foreground extraction of athlete by FCN for sport training video analysis. The extraction by FCN effectively avoids the bias from background.
- CNN is leveraged to estimate frame probability to each pose and extract key frames from athlete training videos with the maximum probability of each pose.
- The proposed deep key frame extraction method successfully applies the deep learning model of FCN and CNN to sport training video analysis.

2 Deep Key Frame Extraction for Sport Training Videos

In this section, we describe the details of the proposed deep key frame extraction (DKFE) method for sport training. Figure 4 illustrates the framework of the proposed method for key frame extraction of sport videos.

In Fig. 4, the proposed DKFE contains two parts: pose probability estimation of each frame and key frame extraction with the strategy of choosing the frame with relatively maximum probability located in the center of similar frames. DKFE focuses on investigating the pose of athlete in sport videos. Therefore, DKFE firstly investigates foreground of the first frame in sport video and employ FCN to extract the foreground region of the athlete and barbell, which effectively avoids the influence from complex background. Then all the frames of the video are cropped referring to the first frame. The CNN model is fine-tuned with training image set and leveraged to estimate key pose probabilities of the

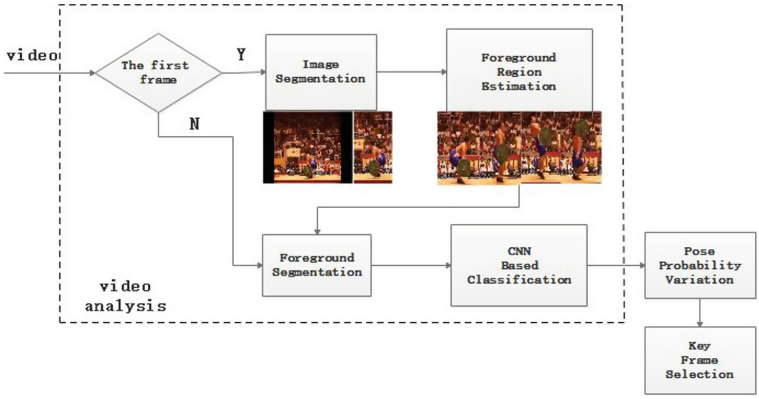


Fig. 4. The framework of the proposed deep key frame extraction method for sport videos.

frames. Finally, DKFE extracts the frames with relatively maximum key pose probabilities.

2.1 Foreground Extraction

In scenario of analyzing sport training videos, the athlete and barbell are considered as foreground region while the audiences, floor and the other objects are treated as background region. We randomly select 373 frames from all the videos labeled foreground and background regions to train a proper FCN model for foreground extraction.

2.2 CNN Based Pose Categorization

Then CNN model is trained over the extracted foreground regions without bias from background to distinct different key poses. We have 3062 key frames from 516 videos totally, where 1837 frames are selected to train a CNN model, 613 frames are collected to validate the model and 612 frames are remained to test the model to category different poses. The proposed DKFE employs CAFFE-NET to category the poses. Moreover, as CAFFE-NET has only 8 layers, it costs a short time to train a proper model. We fine-tune the CNN model in DKFE over the pre-train ImageNet model.

2.3 Pose Probability Estimation

Then the trained CNN model is conducted to estimate key pose probabilities of testing videos. Figure 5 provides the average estimated four key pose probabilities of videos, where x-axis denotes frame sequence of videos and y-axis represents the estimated key pose probabilities. In Fig. 5, four key poses denoted by different colors are investigated and the video plays poses one by one. The results in

Fig. 5 illustrate that every pose focuses on a quarter in order. It implies that CNN based pose probability estimation is able to distinguish frame sequences of the poses with each other. Therefore, we design a key frame extraction strategy based on the CNN probability estimation model.

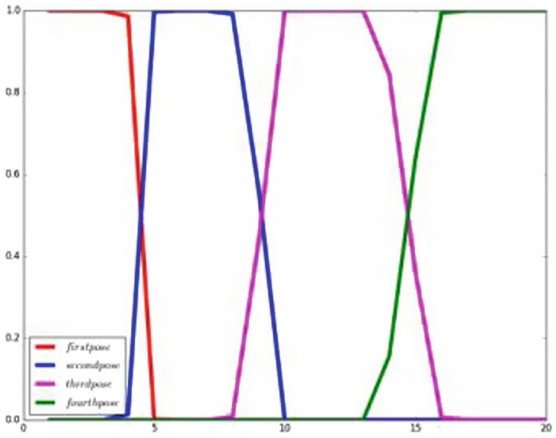


Fig. 5. Average estimated key pose probabilities of frames in sport videos. The four key poses are represented with four different colors, respectively. (Color figure online)

2.4 Key Frame Extraction Strategy

From a sport training video, the target of this work is to select the corresponding frames of key poses. Intuitively we attempt to extract the frame with maximum key pose probability. However as shown in Fig. 5, there exists many frames with the similar key pose probability which means they tend to contain similar foreground contents. Therefore, alternatively we select the center frame as the key frame when the probability difference between the neighboring frames is smaller than a tolerance as

$$F_i = \begin{cases} \arg\max\{p_i^j, i = 1, 2, 3, 4, j = 1, 2, \cdots, N\} & e_i \geqslant E \\ \arg\{p_i = p_i^{M/2}\} & e_i < E \end{cases} \tag{1}$$

where F_i represents the key frame of i-th pose, $\mathrm{P}_i{}^j$ denotes the key pose probability of j-th frame to the i-th pose and M is the number of neighboring frames with similar probabilities. When the probability difference of adjacent frames e_i is larger than E, we select the frame with the maximum key pose probability as the key frame. If the e_i is less than E, we alternatively select the frame of $\mathrm{p}_i{}^{M/2}$ as the corresponding key frame.

3 Experimental Results

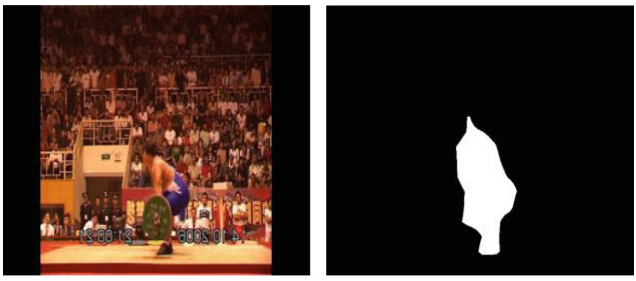
We perform experiments and their corresponding analysis to verify the superiority of the proposed DKFE in key frame extraction of sport videos. In this section, we present the implementation details and compare the performance of the proposed method with some sport key frame extraction methods such as Wu's method [9] and deep learning based method [8]. We conduct experiments on a video database collected from general administration of sport of China, which contains 516 sport videos with various kinds of weightlifting competitions. All the videos are recorded from a side of the athlete.

3.1 Implementation Details

The proposed DKFE framework is built on the deep learning model of FCN [4] and CNN [5]. The whole network is fine-tuned with an initialization of the pre-trained CAFFE-NET. The parameters in FCN based foreground extraction network are set with mini-batch size of 1, base learning rate to 1×10^{-3}, momentum to 0.9, weight decay as 5×10^{-4}, and maximum number of training iterations to 2000. The parameters of CNN based pose probability estimation network are set as mini-batch size of 256, base learning rate to 1×10^{-4}, momentum to 0.9, weight decay to 5×10^{-4}, and maximum number of training iterations to 20000.

3.2 Experimental Analysis

For key frame extraction from sport training videos, the proposed DKFE leverages two deep models of FCN based foreground extraction and CNN based pose probability estimation modules to figure out the frames with key poses. In the following, the performance of foreground extraction with FCN and key pose probability estimation by CNN are presented respectively.



(a) (b)

Fig. 6. An example of foreground extraction result with FCN in DKFE.

Foreground Extraction with FCN. FCN model is trained with randomly selected 373 frames labeled foreground and background regions for athlete related foreground extraction. After a proper FCN model is trained, the corresponding threshold for FCN to separate foreground and background can be determined as 0.85 with exhaustive experiments on randomly selected testing frames. The pixels with the estimated foreground probabilities larger than the threshold are binarized to foreground, vice versa. Figure 6 provides an example frame of FCN based foreground extraction, (b) is the binary mask of foreground extraction by FCN on the frame in (a). The foreground extraction helps effectively prevent bias from the complex background in sport videos.

Key Pose Estimation by CNN. Then a CNN model is trained on the extracted foreground to estimate key pose probabilities of frames. For example, there are four key poses in weight lifting. The trained CNN model is employed to estimate likelihood probability of each frame to the four poses as Fig. 5. As the proportional size of training, validating and testing dataset in 3:1:1, we have 612 frames to test. Table 1 provides the accuracy of each pose estimated by CNN. The results indicate that the trained CNN model successfully estimates the poses of testing frames.

Table 1. Accuracy of each pose by CNN on testing frames.

	Total	Correct	Wrong	Accuracy
Pose1	169	165	4	97.6%
Pose2	130	123	7	94.6%
Pose3	155	152	3	98.1%
Pose4	158	156	2	98.7%

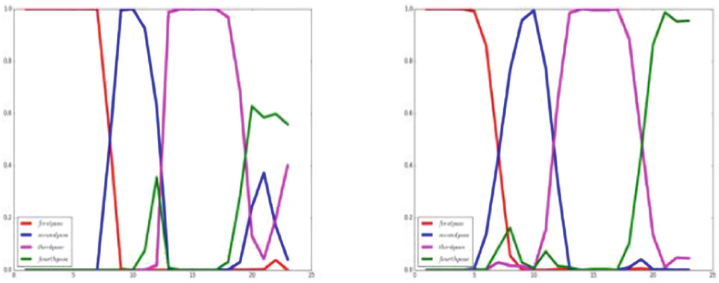
Table 2. Accuracy of DKFE in key frame extraction from sport videos compared with Wus method [9].

Method	Accuracy
Wu [9]	90.6%
DKFE	97.4%

3.3 Key Frame Extraction from Sport Videos

We further compare the proposed DKFE method with Wu's method [9] and deep learning based method [5]. CNN based method [5] did not employ techniques in avoiding bias from complex background. With the same settings to [9], DKFE is compared to [9] to extract key frames of sport videos in Table 2. The results illustrate DKFE outperforms Wus method [9] with almost 7% improvement.

Compared with the CNN method [5], Fig. 7 provides the estimated pose probability of DKFE for key frame extraction. In Fig. 7(a) the fourth pose did not distinct apparently to the others. Because in [5] the feature of the fourth pose is influenced by the complex background and the CNN model could not capture proper characteristics of the pose. However, Fig. 7(b) gives an estimation of the fourth pose as good as that of the other poses. It benefits from the foreground extraction by FCN for CNN based pose estimation for key frame extraction from sport videos. Therefore, we can conclude that the proposed DKFE is capable of extracting key frames from sport videos to aid the sport training.



(a) (b)

Fig. 7. The pose probability estimation of DKFE compared with CNN based method [5] for key frame extraction from sport videos.

4 Conclusion

In this paper, we propose deep key frame extraction (DKFE) for sport training videos. DKFE employs FCN based foreground extraction for the following pose estimation, which effectively prevents bias from complex background to the pose estimation of frames. Moreover, CNN model in DKFE performs to estimate key pose probability of frames in an video and extract the frames of key poses with relatively maximum probability centered at the frames of similar probability. The experimental results demonstrate that the proposed DKFE has a good ability in extracting key frames from sport training videos.

References

1. Ng, J.Y., Hausknecht, M., Vijayanarasimhan, S.: Beyond short snippets: deep networks for video classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4694–4702 (2015)
2. Fan, X., Zheng, K., Lin, Y., Wang, S.: Combining local appearance and holistic view: dual-source deep neural networks for human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1347–1355 (2015)

3. Cheron, G., Laptev, I., Schmid, C.: P-CNN: pose-based CNN features for action recognition. In: Proceedings of International conference on Computer Vision (ICCV), pp. 3218–3226 (2015)
4. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440 (2015)
5. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: ACM Conference on Neural Information Processing Systems (NIPS), pp. 1097–1105 (2012)
6. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8 (2008)
7. Niebles, J.C., Wang, H., Li, F.F.: Unsupervised learning of human action categories using spatial-temporal words. ACM Trans. Int. J. Comput. Vis. **79**(3), 299–318 (2008)
8. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: IEEE Conference on International Conference on Pattern Recognition (ICPR), pp. 32–36 (2004)
9. Wu, L., Zhang, J., Yan, F.: A poselet based key frame searching approach in sports training videos. In: IEEE Conference on Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 1–4 (2012)
10. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1933–1941 (2016)