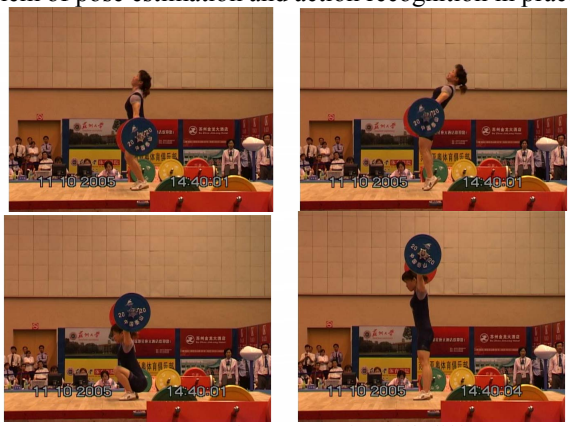
A poselet based key frame searching approach in sports training videos

Lifang Wu* Jingwen Zhang* Fenghui Yan*
*College of Electronic Information and Controll Engineering,
Beijing University of Technology, Beijing, China, 100124.
E-mail:lfwu@bjut.edu.cn Tel: +86-10-67396151

Abstract----In some sport training application, it is necessary to search the key frames of training video for carefully analysis. In this paper, we take the key frame searching issue as a pose estimation problem. First, a set of various pose detectors are collected trough the twice SVM training process, each of which can be interpreted as a learned pose-specific HOG weight classifier. Then we run each linear SVM classifier over the image in a multi_scale scanning mode. In order to resolve the problem of extreme similarity between the adjacent frames, the detection hits at every scale in each frame is counted as the principle of optimal key frame selection. The frame with the most detection hits are chosen as the key frame for the pose detector. The experimental results using weight-lifting training videos show the efficiency of proposed approach.

I. INTRODUCTION

With the development of sports, the competition is becoming more and more intensive. Scientific training is one of key issues. In some sports such as track and field, fencing, diving, gymnastics and so on, accurate pose analysis is important for effective training. Let's take weight lifting for example, it includes four phases: First, the athletes lift the bell and extend their knees. Then the force is burst out and the athletes lift the bell fast. Next, the athletes squat down and lift the bell over their head. Finally, the athletes stand and keep up the bell. The corresponding key poses of these phases are shown in Fig. 1. The joints of these key poses should be located and these poses of different athletes should be analyzed. These objective data is helpful for training and it requires searching the corresponding key frames firstly. This problem is superficially key frame searching, but it is a problem of pose estimation and action recognition in practice.



Figue1 Four key postures in weight lifting

Some researchers estimated actions from videos [7, 8, 9] using motion cues. But the extremely inter-frame similarity in some videos usually caused unsuccessful estimation. Some researchers proposed approaches based on the pictorial structures framework [4, 5, 12]. The strong discriminatively appearance models were trained to localize the spatial layout of the human body. Then a feature vector or a flexible kinematic tree is used to recognize the action.

Recently, Felzenszwalb et al. [11] thought that part_based representations could capture the pose variations of an object effectively. Furthermore, Bourdev and Malik proposed poselet [2, 3] for human body detection, object segmentation and pose estimation. They constructed a dataset of Human bodys for selecting good poselets. The dataset is unique in cross-referencing the full 3D pose, keypoint visibility and region annotations. They used weighted least squares fitting to find the poselet candidates. Then they extracted the HOG [1] of these candidates. Finally, they obtained the poselets SVM classifies using HOGs of these candidates. In the detecting stage, they proposed a two-layer classification model for detecting people and localizing body components. Yang et al [10] also utilized the poselet for pose estimation. They presented a model that integrated action recognition and pose estimation.

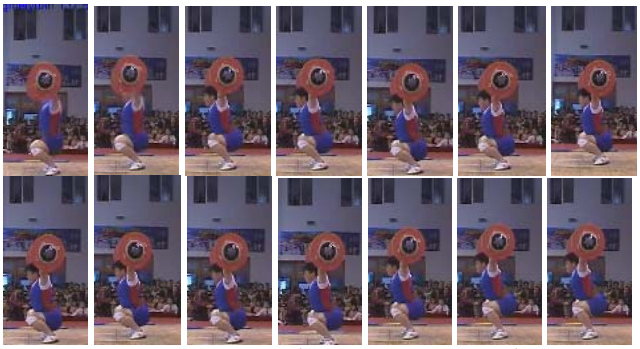


Figure2 some frames of 75th frame through 102th frame in a training video

In our application scenario, there are two difficulties in our problem: 1) The athlete's body is always partially occluded by the Barbell, which make it difficult to get the articulated pose configuration. Therefore, the pictorial structures approach doesn't work. 2) There is much inter-frame similarity in the video. Fig2 shows some frames in the third phase from a video. In the total 27 frames, the athlete has almost similar pose. There is minimal change in the pose of these

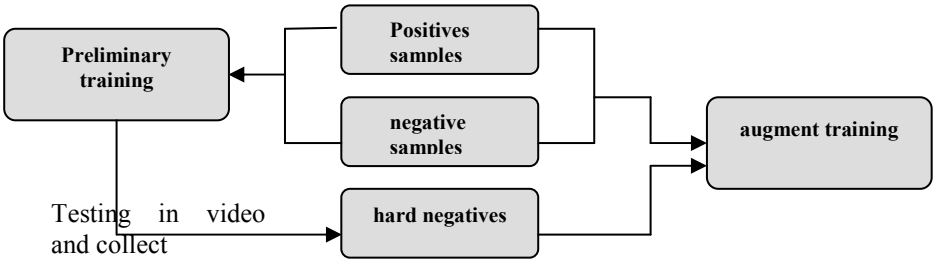


Fig.3 Training stage of the proposed approach

frames. It is possible that the motion cue based approaches do not work in such videos.

The part-based approach attempts to address such problem by manually labeling parts and using them to train a set of pose-specific detectors. While very encouraging, the heavy manual labeling burden is a big limitation of this method. Furthermore, visual correspondence between the learned model and the detected instance is very coarse with these part-based poselets. How could we find the most similar posture we need in a series of successive frames? In this paper, we propose a statistic-based scheme to resolve these problems.

The remaining parts of this paper are organized as follows: In Section 2, we describe the proposed scheme. We introduce the framework of proposed scheme. We get the final classifier by preliminary training and augment training. Then we search the key frame using the statistics of detection hits. In Section 3, we illustrate the experimental results as well as the analysis of these results. Finally, Section 4 concludes this paper with a summary

II. THE PROPOSED APPROACH

Our poselet detector is based on a very simple idea of training a independent classifier for each posture. Just like Navneet Dalal and Bill Triggs, we represent each posture using a rigid HOG template. In our work, not the partial but the whole human body is utilized to represent a poselet. We use the linear SVM classifiers, and each classifier can be interpreted as a learned pose-specific HOG weight vector. As a result, instead of lots of complex part-based detector, we have a collection of simpler individual SVM detectors of various postures, each highly tuned to the appearance which we want to get. One poselet we defined is as shown in Fig.4.

The framework of the proposed approach is shown in Fig.3. First we extract the HOG of both positive and negatives samples. Then the preliminary training is implemented and a preliminary classifier is obtained. Then all the frames of training video samples are tested using the preliminary classifier. The mis-classified patches are used as hard-negative sample for augment training and we get the augment classifier. In the posture searching stage, for each frame in a video, the specified pose is detected using augment classifiers in multi-scale scanning mode. Unlike the usual practice that gives the location of the object by the clustering, the number of detection hits at every scale is counted in each frame to find the most similar posture. The frame with the most detection hits is chosen as the key frame for this posture. （the specified poselet classifiers are used to detect our postures in order）

2.1 The preliminary training

In a specific kind of sports training videos, a fewer key actions of athletes are needed. Therefore, unlike Bourdev and Malik[1], we don't need to do a lot of manually labeling. We just need to scaled tailor the whole body under a specific posture in each frame with appearance similarity. Some samples of a poselet are shown in Fig.4. These samples and their mirror edition are utilized as the positive samples in training stage.

The negative samples are collected by cutting the training frames randomly. It is certain that the negative should not include the whole human body. Some of the negative samples are shown in Fig.5.

Fig.4. some samples of a poselet

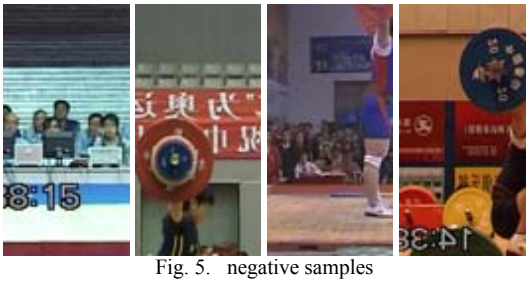


Fig. 5. negative samples

Although the number of positive samples is relatively smaller than that of negative samples, they are tightly clustered due to the consistency of configuration and appearance. We need to increase the number of negative samples, so that the decision boundary of the class depends on not only the positive sample (“what it is”) but also the negative samples (“what it is not”) Ref.[6]

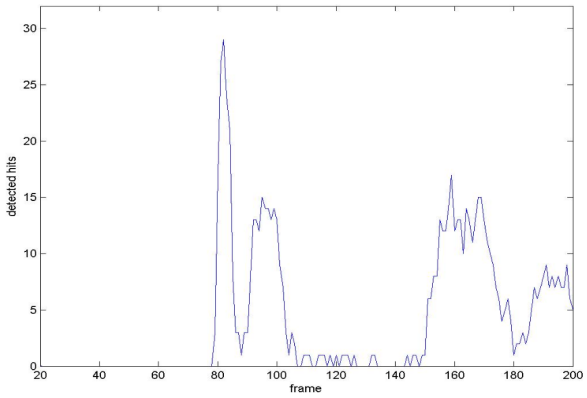


Fig. 6 The detection hits of the second poselet in each frame of a video using the preliminary SVM classifier

Fig. 7 show that the preliminary SVM classifier is able to detect the specified pose, but there are a lot of false detections which are far from the appearance we want. Fig 6 shows the detection hits of the second poselet in each frame in a video using the preliminary SVM classifier. Such a classifier is unsafe for it will cause a lot of false detection, and the efficiency of such a classifier is very low. Therefore, the augment training is needed.

2.2 The augment training

In this step, the mis-classified samples are adding to the negative set for augment training.

What we have to mention is that these detected patches/windows which are mis-classified while very similar to the posture should not be added into the hard-negative samples. It will result in a sensitive classifier which can't recognize correct object. Furthermore, owing to the principle of optimal key frame selection in the third step, we don't need to require the detect results in strict conformity with the current poselet classifier.

Fig. 7 Some patches of the second poselet in each frame of a video using the preliminary SVM classifier

The augment classifier is tested in the same video. The number of detection hits at each frame is shown in Fig. 8. We can see that the detection hits is more clustering. It will make the following process easier.

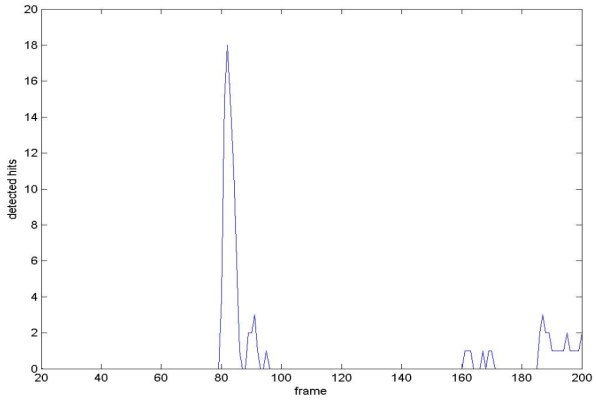


Fig. 8. The hits of in each frame of the testing video same as in Fig. 6 using the augment SVM classifier

Although the hits are more clustered, we also could find the detection hits in more frames. How could we determine the best frame? We use the statistics of detection hits.

2.3 key frames search using statistics of detection hits

Suppose that there are N frames in a video. For the frames, we count the detection hits number h_i^1 (the sum of detection hits at every scale in a frame) of the first classifier in all the N frames.

Then we find the frame (m_1) which is detected with the maximum hit number.

$$m_1 = \arg\max\{h_i^1, i = 1,2,...N\} \quad (1)$$

For the second posture, we count the detection hits number h_i^2 of the second classifier in the rest $N - m_1$ frames.

And the frame (m_2) detected by the second classifier with the maximum hit number is taken as the second key posture..

$$m_2 = \arg\max\{h_i^2, i = m1+1,...N\} \quad (2)$$

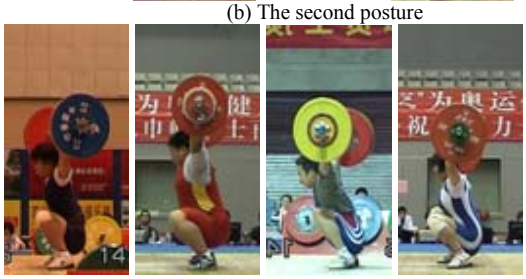
By the similar way, we could search the frames which include the third and forth key postures.

III. EXPERIMENTAL RESULTS

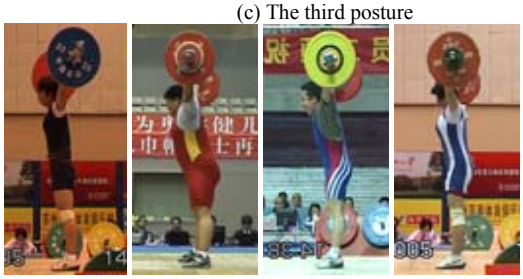
We use the proposed approach to search four key postures in the weight lifting video. 22 videos are used to collect the training samples. Generally, 20-30 samples for each pose are collected from the training videos. Some samples are shown in Figure 4. And we deal these samples with the left_right reflections. We get total 40-60 positive samples. Then we randomly segment a set of 300-400 patches in the training videos as the initial negatives samples.

(a) The first posture

(b) The second posture



(c) The third posture



(d) The forth posture

Fig.9 our experiment four poselet

We test the approach using 22 training videos and 8 testing videos. The experimental results are shown in Table 1.

Table 1 the posture search results

		Total	Detect	percentage
Training set	Pose1	22	22	100%
	Pose2	22	22	100%
	Pose3	22	22	100%
	Pose4	22	22	100%
	Total	88	88	100%
Testing set	Pose1	8	8	100%
	Pose2	8	8	100%
	Pose3	8	6	75%
	Pose4	8	7	87.5%
	Total	32	29	90.6%

From Table 1 we can see that our approach could detect 90.6% posture for the testing set. Fig.10 show some detection results.

Figure 10 Some posture search results

IV. CONCLUSION

In this paper we propose a poselet based key frame searching approach for sport training. We define 4 postures. For each posture, the whole human body is collected as positive samples. Using these samples, a collection of simpler individual SVM detectors is trained. The detection hits of each frame are taken as the selection criteria of key frame for the specified pose. The experimental results show that the proposed approach could search over 90% postures correctly.

ACKNOWLEDGMENT

This paper is supported by the program of Beijing Municipality excellent under Grant No 2009D005015000010.

REFERENCES

[1] N. Dalal and B. Triggs. “Histograms of oriented gradients for human detection.” .In CVPR, 2005.

[2] L. Bourdev and J. Malik. “Poselets: body part detectors training using 3D human pose annotations.”. In ICCV, 2009.

[3] L. Bourdev, S. Maji, T. Brox, and J. Malik. “Detecting people using mutually consistent poselet activations”. In ECCV, 2010.

[4] M. Andriluka, S. Roth, and B. Schiele. “Pictorial structures
revisited: People detection and articulated pose estimation.” In CVPR 2009.

[5] Vajda, T, Zoltán, A “Pictorial Structure Based People Detection and Pose Estimation in Videos” .IEEE 2011

[6] T. Malisiewicz. A. Gupta. Alexei A. Efros “Ensemble of Exemplar-SVMs for Object Detection and Beyond”. ICCV 2009

[7] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. “Learning realistic human actions from movies”. In CVPR, 2008.

[8] J. C. Niebles, H. Wang, and L. Fei-Fei.“ Unsupervised learning ofhuman action categories using spatial-temporal words”. In BMVC, volume 3, pages 1249–1258, 2006.

[9] C. Schuldt, I. Laptev, and B. Caputo. “Recognizing human actions: a local SVM approach”. In ICPR,2004

[10] Weilong Yang, Yang Wang, and Greg Mori “Recognizing Human Actions from Still Images with Latent Poses”. in CVPR,2010

[11] P. Felzenszwalb, D. McAllester, and D. Ramanan. “A discriminatively trained, multiscale, deformable part model”. In CVPR, 2008

[12] V. Ferrari, M. Mar´ın-Jim´enez, and A. Zisserman. “Pose search: retrieving people using their pose”. In CVPR, 2009