ACTION RECOGNITION IN UNCONSTRAINED AMATEUR VIDEOS

Jingen Liu[1], Jiebo Luo[2], Mubarak Shah[1]

[1]Computer Vision Lab, University of Central Florida [2] Eastman Kodak Company

ABSTRACT

In this paper, we propose a systematic framework for action recognition in unconstrained amateur videos. Inspired by the success of local features used in object and pose recognition, we extract local static features from the sampled frames to capture local pose shape and appearance. In addition, we extract spatiotemporal features (ST features), which have been successfully used in action recognition, to capture the local motions. In the action recognition phase, we use the Pyramid Match Kernel based on weighted similarities of multi-resolution histograms to match two videos within the same feature types. In order to handle complementary but heterogeneous features, i.e., static and motion features, we chose a multi-kernel classifier for feature fusion. To reduce the noise introduced by the background clutter, our system also tries to automatically find the rough region of interest/action. Preliminary tests on the KTH action dataset, UCF sports dataset, and a YouTube action dataset have shown promising results.

Index Terms— Action Recognition, Video Analysis, Video Indexing

1. INTRODUCTION

With the explosive proliferation of digital video in people's daily life and on the Internet, action recognition is receiving increasing attention due to its wide range of applications such as video indexing and retrieval, activity monitoring in surveillance scenarios, and human-computer interaction. Most of the earlier research work focused on holistic video representations such as spatiotemporal volume [3] or trajectories of body joints [4,5]. In order to obtain reliable features, these approaches often make certain strong assumptions about the video. For instance, the systems in [4,5] require reliable human body joint tracking, and [3] needs to perform background subtraction to create a 3D shape volume. Although the holistic methods have obtained high recognition accuracy on simple video sequences taken in carefully-controlled environments with largely uncluttered background, and without camera motion, these strong assumptions limit their application to a more complicated dataset than the commonly used "clean" KTH dataset [8]. In real practice, it is simply not feasible to annotate a large video dataset to obtain body joints, or perform reliable background subtraction on a dataset that often contains significant camera motion.

Fig.1. Examples of *Horseback riding* and *Soccer juggling* under different viewpoints, scales and illuminations.

Inspired by the success of local features in object and scene recognition [6,7], bag-of-video-words (BOW) approach has been exploited recently in action recognition [8,9,10]. Compared to the holistic approaches, it does not require background subtraction and tracking, and it can cope with small camera motion and illumination changes. Typically, spatiotemporal interest points are first detected either by a 3D Harris corner detector [11] or 1-D Gabor filters [12], and they are then quantized into video-words whose statistical distributions are used to represent the entire video sequence. Once an action video is represented by BOW, we can either choose a discriminative learning model such as SVM or a generative model such as pLSA to build a classifier. More attempts based on BOW have been made and impressive results were obtained on the KTH dataset [8, 9, 10]. However, very few attempts have been made on amateur videos as shown by the examples in Fig. 1. Here, amateur video refers to a video taken under uncontrolled capture conditions, such as by a hand-held camera. Because of the diverse video sources, including YouTube, TV broadcast and personal videos, amateur videos generally contain large variations in camera motion, background clutter, viewpoint change, illumination condition, object appearance and scale, and other aspects. Laptev et. al. [19] reported promising results on action recognition from movie clips, which are still not as challenging as amateur videos.

On the other hand, the human vision system seems to be capable of recognizing many types of human actions from an instantaneous pose in a single image without motion information. In computer vision, pose recognition from local shape features such as shape context [13], appearance and position context [14] have also obtained good performance for action recognition. The advantages of action recognition from a single pose are at least two-folds: action recognition can then be treated as a special case of object recognition,

and this is desirable when motion features are unreliable (e.g., due to unpredictable camera motion).

Therefore, we can perform action recognition using either local static features or local motion features. It is difficult, however, to argue which is consistently better. In fact, they are complementary. For instance, suppose we want to differentiate *Cycling* from *Horseback riding*. Our observation is that both actions cause similar camera motion such as panning. It is difficult to distinguish them by motion features because not only the horizontal motion in these two actions are similar, but the typical cluttered background would also introduce a large amount of noise into the motion features for in both cases. Yet, we can easily tell *bicycles* from *horses* based on their local shapes or appearance features. In this case, static features may outperform motion features. For another example, it is conceivable that we cannot identify *jogging* from *running* only based on a static pose and must use motion features to make the distinction. Therefore, we propose to incorporate both static features (local shape and appearance) and spatiotemporal motion features (local motion).

Little work has been reported on using a combination of static and motion features for action recognition in amateur videos until recently. Neibles et al. [15] proposed a generative model to use both static and dynamic features for action recognition on a simple dataset. It is difficult to extend this model to a large video dataset, and in particular amateur videos due to their complexity. Instead of detecting expensive spatiotemporal interest points, Mikolajczyk et al. [16] only extract static features with associated motion vectors from each frame. Motion vectors are used as a filter in recognition. Their action recognition method is akin to object recognition, and also requires extra training images and bounding boxes for training. Ignoring the noise introduced by different features, Schindler et al. [17] combine different types of ST features by simply concatenating the feature vectors.

In this paper, we prose a systematic framework for action recognition in amateur videos. In addition to ST features that capture local motion, we also extract static interest points from temporally sub-sampled frames, which contain the local shape or appearance information. Unlike the previous work that combined the static and motion features by simply concatenating them, our system is able to automatically find the discriminative static features. This is motivated by our observation on amateur videos that the background is highly variable; simply combining the static and motion features may introduce noise in the features. Instead, it is beneficial to first separate discriminative foreground features from the cluttered background. Furthermore, in order to effectively select the good features from the static and motion features, we investigate the use of multi-kernel classifier to combine such heterogeneous features [2,18]. Within the same type of features, we use the Pyramid Math Kernel [1] to obtain better action matching.

2. VISUAL FEATURE EXTRACTION

This section discusses the static and motion features:

Static features. Local features have been widely used for object recognition and scene classification due to their scale, view, rotation and translation invariance. For every sampled frame, we first apply three interest point detectors: Harris-Laplacian (HAR), Hessian-Laplacian (HES) and MSER detectors. The three detectors can produce complementary features. HAR locates corner features in an image, and both HES and MSER extract blob features that are complementary to corner features. Next, each feature is described by their location (x, y), their scale σ and a 128-dimensional SIFT descriptor.

Motion features. We adopt the spatiotemporal interest point detector proposed by Dollar et. al. [11]. Compared to the 3D Harris-Corner detector, it produces dense features that can significantly improve the recognition performance in most cases. We apply two separate filters in spatial and temporal directions. For each pixel, a response value is given by the following function,

$$R = \{I * g_\sigma(x,y) * h_{ev}(\mathrm{t})\}^2 + \{I * g_\sigma(x,y) * h_{od}(\mathrm{t})\}^2,$$

where $g_\sigma(x,y)$ is the spatial Gaussian filter, h_{ev} and h_{od} are a quadrature pair of the 1-D Gabor filter in temporal direction. This detector produces high responses to temporal intensity change points. The interest points are selected at the locations of local maximal responses, and 3D cuboids are extracted around the interest points. For simplify, we use the flat gradient vectors to describe the cuboids. Furthermore, PCA is utilized to reduce the descriptor dimension to a smaller number (e.g., 100 in this study).

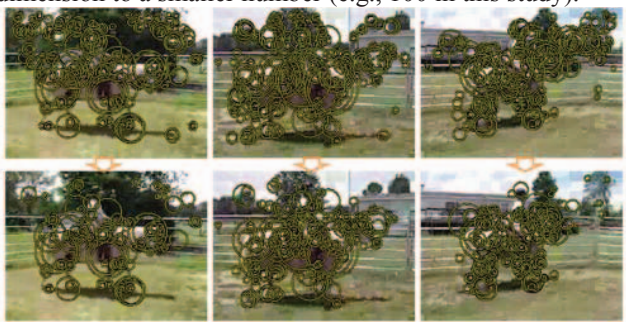

Fig. 2 Effects of finding ROI. First row shows all static features; second row shows the selected features in ROI.

3. REGION OF INTERESTS

We introduce static features into our action recognition framework because the pose is one of the important cues for action recognition. We also noticed that even though amateur videos usually contain cluttered background, the background can provide useful contextual information for recognition. For example, football field can help detect football-related actions. Nevertheless, the background can vary dramatically even for the same type of actions. Therefore, it is always helpful to locate the region of interest in the video. We believe that the local shape or appearance information in the region of interest, combined with (weak) contextual information from the background, provides the

best opportunity for action recognition in amateur videos. In order to design the best strategy for combining the heterogeneous sources of information, we chose to consider amateur videos in two categories: A) videos with relatively fixed background, and B) videos with moving background (e.g., horseback riding and cycling).

Type A. For this type of videos, we are interested in the area of high motion. The shape and appearance features extracted from these areas are more important. The computation of ROI is straightforward. Suppose the current frame is F_t, and $W = \{w_i(x_i, y_i, t_i, d_i) | 1 \le i \le n, t - \sigma \le t_i \le t + \sigma\}$ is a set of spatiotemporal cuboids detected around the current frame in the time span of 2σ (e.g. . $\sigma = 4$ in this paper). We can estimate the centroid of the region as $\hat{x} = \frac{1}{n}\sum_i x_i$, $\hat{y} = \frac{1}{n}\sum_i y_i$, and the dimensions of the region are $D_x = 2\sqrt{3c_{xx}}$, $D_y = 2\sqrt{3c_{yy}}$, where c_{xx} and c_{yy} are the central moments with respect to the centroid.

Type B. This type of videos is taken by moving cameras. Because the background is constantly changing, the appearance features of the background are not persistent through the sequence. Meanwhile, the appearance features for the foreground (ROI) can be persistently tracked throughout the sequence. For simplicity, we use histogram matching to find the persistent appearance features, based on the bag-of -visual-words image representation. We apply k-means clustering to the appearance features of the entire video, and obtain a K-entry codebook $C = \{c_i | 1 \le i \le K\}$. Next, each frame is represented by a histogram of the visual words present in that frame. For each visual word c_i we compute its mean value and covariance values,

$$\bar{h}(c_i) = \frac{1}{M}\sum_{j=1}^{M} h_j(c_i), \quad \sigma(c_i) = \frac{1}{M}\sum_{j=1}^{M} (h_j(c_i) - \bar{h}(c_i))^2$$

where M is the total number of frames, $h_j(c_i)$ is the histogram of visual word c_i for frame j. Small σ value means that the visual word is more persistent in the video. We consider a visual word c_i is a good feature for this video if it satisfies $\bar{h}(c_i) \epsilon (h_{LB}, h_{HB})$ and $\sigma(c_i) \le \varepsilon$. The examples in Fig. 2 demonstrate the effect of finding ROI.

4. MULTI-KERNEL CLASSIFIER

Kernel-based classifiers such as SVM have been widely used and the kernel selection is very important for good performance. In this section, we discuss how to fuse homogenous features (of the same type) using the Pyramid Mach Kernels, and heterogeneous features (of different types) using linear combination of kernels.

4.1. Pyramid Match Kernels (PMKs)

For each type of feature f (i.e., ST features, HAR, HES, MSER), we randomly select a set of features F from the training videos. We then use the hierarchical k-means algorithm to cluster them with L+1 levels. If we consider each level except for the first one as a vocabulary, we obtain L visual vocabularies $V_f^1, \dots, V_f^L$, where V_f^1 is the coarsest one and V_f^L is the finest one. Given an action video S, we compute a histogram $H_f^l(S)$ based on a vocabulary V_f^l.

In order to match two videos, we use PMKs. The similarity between two videos S_p and S_q is computed as follows,

$$K_f(S_p, S_q) = \sum_{l=1}^{l=L} \rho^l \exp\left(-\frac{d(H_f^l(S_p), H_f^l(S_p))}{\mu^l}\right)$$

where ρ^l is the weight assigned to matching made at level l (the weight can be either learnt from a validation dataset or predefined as $\rho^l = 2^{-(L-l)}$), μ^l is a scaling parameter that can be set to the mean value of d distance between the l level histograms over the training videos, $d(H_f^l(S_p), H_f^l(S_p))$ is the distance measure between two histograms. Instead of using the histogram intersection kernel, we adopt a chi-square kernel which produces superior performance according to previous research [18].

4.2. Dynamic Combination of Heterogeneous Features

In order to combine the video matching results of all the heterogeneous features, we introduce the following dynamic linear kernel combination to obtain the final classifier kernel,

$$K(S_p, S_q) = \sum\nolimits_f \gamma_f K_f(S_p, S_q)$$

where γ_f o represents the weight assigned to ST features, HAR, HES, or MSER features. The weights are optimized by minimizing the classification rate over the validation dataset ($ERR(Va)$ is the classification error on validation dataset):

$$\gamma_f = \underset{\gamma_f}{arg\min}\, ERR(Va)$$

5. EXPERIMENTS

5.1. KTH Dataset

The KTH dataset is the most popular dataset for action recognition in a controlled, "clean" environment. It contains 6 actions performed by 25 actors in four different scenarios, for a total of 598 video sequences. We extract at least 200 cuboids from each video, and then apply hierarchical k-means to generate (L=3) visual vocabularies (The size of the vocabulary is 20, 200 and 1000 for each level, respectively). As for the static features, we uniformly sample 40% of frames in a video, and from each frame extract 5 to 30 interest points for each type of static feature. Similarly, hierarchical k-means is used to create L = 3 visual vocabularies with size of 20, 200 and 1000 for each level, respectively. We adopt a 5-fold cross validation scheme. For the multi-kernel combination, we learn the weights for HAR, HES, MSER and ST feature, and they are 0.2, 0.1, 0.1, and 0.6 respectively.

Fig. 3 shows the classification results using motion features, static features and the combination of both. In terms of average accuracy, motion features outperform static features, although static features did much better than motion features for *boxing*. The combination of both features led to 6.0% improvement over using motion

features alone. The average accuracy is better or competitive to 80.7%[11], 91.8%[19], 94.1%[8] and 91.6%[10].

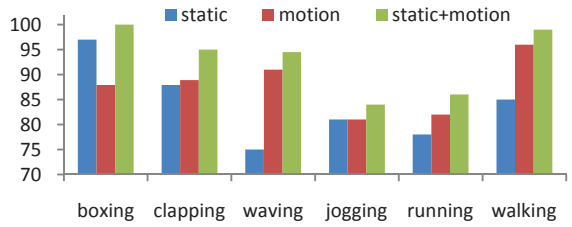

Fig. 3 Recognition performance on KTH dataset. The average accuracy for static, motion and static+motion experimental strategy is 87.3%, 83.7% and 93.1% respectively.

5.2. UCF sports Dataset

UCF sports dataset contains 10 sports actions: *diving*, *golf swing*, *kicking*, *lifting*, *horseback riding*, *running*, *skateboarding* (sk_b), *swing angle* (sw_a), *swing bench* (sw_b) and *walking*, with a total of 150 video sequences at the resolution of 720x480. It is a very challenge dataset due to the camera motion and background clutter. Half of the videos are used for training, and the rest for testing. We repeat it for 10 times. Finally, average performance for each action is reported in Fig. 4. The experimental results further verify that static and motion features are complementary for action recognition, and the fusion of them can significantly improve the accuracy (e.g. by about 5% for this dataset). Ref. [2] reported an average accuracy of 69.2% on this dataset, using a holistic model with optical flow features.

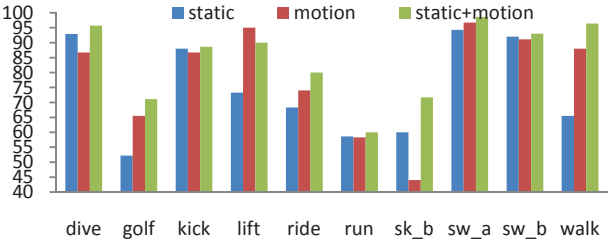

Fig. 4 Recognition performance on UCF sports dataset. The average accuracy for static, motion and static+motion experimental strategy is 74.5%, 79.6% and 84.5% respectively.

5.3. YouTube Dataset

UCF sports videos are still not quite representative of amateur videos, so we collect six actions (i.e. *cycling*, *golf swing :G_swing*, *horseback riding*, *soccer juggling*, *tennis swing* :*T_swing* and *spiking*) from YouTube videos. Each action has 25 independent groups of videos, where different groups are either taken in different environments or by different photographers. With each group having 4 or more videos, there are more than 600 sequences. This dataset is more challenging than the UCF sports dataset because not only it is a much larger dataset, but the videos also have much lower resolution and more cluttered background. We use the same experiment setup as on the KTH dataset, and report the results in Fig. 5. The result verifies our earlier conjecture that *cycling* and *riding* have similar motion features so they are better differentiated by static features.

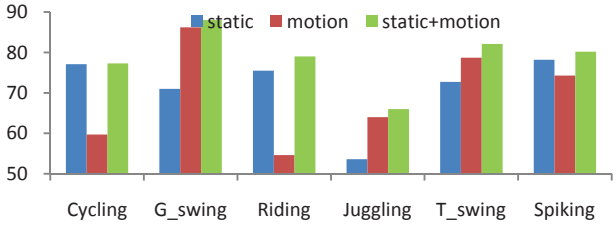

Fig. 5 Recognition performance on YouTube dataset. The average accuracy for static, motion and static+motion experimental strategy is 71.4%, 69.6% and 78.1% respectively.

6. CONCLUSION

For the three video datasets with increasing difficulty, motion and static features complement each other to produce improved action recognition. If more precise object tracker is available, we can further improve the system. Besides, our work is extendable to deal with the scenario where multiple types of actions exist simultaneously.

7. REFERENCES

[1] K. Grauman and T. Darrell. The pyramid match kernel: discriminative classification with sets of image features, *In ICCV 2005*.
[2] M.. Sullivan and M. Shah. Action MACH: Maximum Average Correlation Height filter for action recognition, *In CVPR 2008*.
[3] A. Yilmaz and M. Shah. Action sketch: a novel action representation, *In IEEE CVPR 2005*.
[4] V. Parameswaran and R. Chellappa. Human action recognition using mutual invariants, *In CVIU, 98(2),2005*.
[5] S. Ali and M. Shah. Chaotic invariants for human action recognition, *In ICCV 2007*.
[6] F. Jurie and B. Triggs . Creating efficient codebooks for visual recognition, *In ICCV 2005*.
[7] J. Sivic, B. Russell, A. Efros, A. Zisserman and W. Freeman. Discovering objects and their location in Images, *In ICCV 2005*.
[8] J. Liu and M. Shah. Learning human action via information maximization, *In CVPR 2008*.
[9] J. Liu and M. Shah. Recognizing human actions using multiple features, *In CVPR 2008*.
[10] S. Wong, T. Kim and R. Cipolla . Learning motion categories using both semantics and structural information, *In CVPR 2007*.
[11] P. Dollar, V. Rabaud, G. Cottrell and S. Belongie. Behavior recognition via sparse spatio-temporal features, *In VS-PETS 2005*.
[12] I. Laptev and T. Lindeberg. Space-time interest points, *In ICCV 2003*.
[13] Y. Wang, H. Jiang, M.S. Drew, Z. Li and G. Mori. Unsupervised discovery of action classes, *In CVPR 2006*.
[14] H. Ning, Y. Hu and T.S. Huang . Discriminative learning of visual words for 3D human pose estimation, *In CVPR 2008*.
[15] J. Niebles and L. Fei-Fei. A hierarchical model of shape and appearance for human action classification. *In CVPR 2007*.
[16] K. Mikolajczyk and H. Uemura. Action recognition with motion-appearance vocabulary forest, *In CVPR 2008*.
[17] G. Schindler, L. Zitnick and M. Brown. Internet video category recognition, *In IEEE workshop In Internet Vision 2008*.
[18] A. Bosch,A. Zisserman and X. Munoz. Representing shape with a spatial pyramid kernel, *In CIVR 2007*.
[19] I. Laptev,M. Marszalek, C. Schmid and B. Rozenfeld. Learning realistic human actions from movies, *In CVPR 2008*.