ACTION DETECTION
USING MULTIPLE SPATIAL-TEMPORAL INTEREST POINT FEATURES

Liangliang Cao[+*], *YingLi Tian*[♮], *Zicheng Liu*[♯], *Benjamin Yao*[‡], *Zhengyou Zhang*[♯] *and Thomas S. Huang*[+]

[+] Beckman Institute and Coordinate Science Lab, Dept. ECE, University of Illinois at Urbana-Champaign
[♮] Department of Electrical Engineering, City College of New York
[‡] Department of Statistics, University of California, Los Angeles
[♯] Communication and Collaboration Systems Group, Microsoft Research, Redmond

ABSTRACT

This paper considers the problem of detecting actions from cluttered videos. Compared with the classical action recognition problem, this paper aims to estimate not only the scene category of a given video sequence, but also the spatial-temporal locations of the action instances. In recent years, many feature extraction schemes have been designed to describe various aspects of actions. However, due to the difficulty of action detection, *e.g.*, the cluttered background and potential occlusions, a single type of features cannot solve the action detection problems perfectly in cluttered videos. In this paper, we attack the detection problem by combining multiple Spatial-Temporal Interest Point (STIP) features, which detect salient patches in the video domain, and describe these patches by feature of local regions. The difficulty of combining multiple STIP features for action detection is two folds: First, the number of salient patches detected by different STIP methods varies across different salient patches. How to combine such features is not considered by existing fusion methods [13] [5]. Second, the detection in the videos should be efficient, which excludes many slow machine learning algorithms. To handle these two difficulties, we propose a new approach which combines Gaussian Mixture Model with Branch-and-Bound search to efficiently locate the action of interest. We build a new challenging dataset for our action detection task, and our algorithm obtains impressive results. On classical KTH dataset, our method outperforms the state-of-the-art methods.

1. INTRODUCTION

In the past few years, computer vision researchers have witnessed a surge of interest in human action analysis through videos. Human action recognition was first studied under well controlled laboratory scenarios, e.g., with clean background and no occlusions [18]. Later research work shows that action recognition is important for analyzing and organizing online videos [14]. Moreover, action recognition plays a crucial role in building surveillance system [7] and studying customer behaviors. With the increasing of web video clips and the surveillance systems, it has become very important to effectively analyze video actions.

Fig. 1. Comparing the differences between action classification and detection. (a): for a classification task we need only estimate the category label for a given video. (b) for an action detection task we need not only estimate the category of the action but also the location of the action instance. The blue bounding box illustrate a desirable detection. It can be seen that the action detection task is crucial when there is cluttered background and multiple persons in the scene.

An effective analysis of video actions requires that the systems can answer not only "which action happens in the video", but also "when and where the action happens in the video sequences". In other words, it is preferred to detect the action locations in the videos than simply classifying the video clip to one of the existing labels. When the video file is very long or contains multiple action, simple classification results are not useful. In practice, a surveillance video can be as long as several hours, and a Youtube video might contains quite a few different actions, where only the action detection results algorithm can provide meaningful results.

Despite its importance, action detection is known to be a challenging task. In complex scenes, the background is often cluttered, and the crowds might occlude each other. In this case, it is difficult to distinguish the interesting action with other video contents. The appearance of the actor might look similar as the background. The motion field of the ac-

*Cao would like to thank the support from UIUC CSE fellowship.

tion might be blocked by the other people in the scene. Due to the difficulty of locating the human action, most famous human action data sets [18] [1] involve only the classification task but not location, where the human actions are usually recorded with clean backgrounds, and each video clip mostly involves only one type of action (e.g., running or jogging) and only one person, who keeps doing this action within the whole video clip.

This paper considers the action detection problem using multiple STIP features [6] [11] [19]. An action is often associated with multiple visual measurements, which can be either appearance features (e.g., color, edge histogram) or motion features (e.g., optical flow, motion history). Different features describe different aspects of the visual characteristics and demand different metrics. How to handle heterogeneous features for action detection becomes an important problem.

The difficulty of combining multiple features lies in the heterogeneous nature of different features. Different STIP features are based on different detectors, and the number of detected features varies significantly. It is still an open question how to effectively combine such features. A naive approach is to quantize STIP features and build histogram based on quantization indices. However, much information is lost in the quantization process, and a histogram representation overlooks the differences in the number of detected features. As a result, simply combining histograms will result a poor detection results. This work employs a probabilistic representation of the different features so that we can quantitatively evaluate the contribution from different features. In our approach, we model each feature vector with a Gaussian Mixture Models (GMMs). GMMs with large number of components is known to have the ability to model any given probability distribution function. Based on GMMs, we can estimate the likelihood of each feature vector belonging to a given action of interests. The likelihood can be viewed as normalized contribution from different features, and the optimal bounding box corresponds the maximum of likelihood. The bounding box is found by a branch-bound search [10], which is shown to be efficient and effective to locate the action of interest.

2. RELATED WORKS

Motivated by the recent success of SIFT and HOG in image domain, many researchers have designed various counterparts to describe the spatial salient patches in video domain. Laptev and Lindeberg [11] generalized Harris detector to spatial-temporal space. They aim to detect image patches with significant local variations in both space and time. and compute their scale-invariant spatio-temporal descriptors. This approach is later improved by [12] which gives up scale selection but uses a multi-scale approach and extract features at multiple levels of spatio-temporal scales. The improved method yields reduced computational complexity, denser sampling, and suffers less from scale selection artifacts. Another important video feature is designed by Dollar *et al.* [6], which detects the salient patches by finding the maximum of temporal Gabor filter responses. This method aims to detect regions with spatially distinguishing characteristics undergoing a complex motion. In contrast, patches undergoing pure translational motion, or patches without spatially distinguishing features will in general not induce a response. After the salient patches are detected, the histogram of 3D cuboid is introduced to describe the patch feature.

Many action classification systems [11], [6], [15] [8], [21], [22] are built using Laptev's or Dollar's features. These two features focus the short-term motion information instead of long-term motion, and motion field of a salient patch sometime is contaminated by the background motions. However, most of existing systems only classify the video clips to one of predefined categories, and does not consider the location task.

To overcome the limitation of existing salient patches descriptors, a hierarchical filtered motion field method has been proposed recently for action recognition [19]. This work applies global spatial motion smoothing filter to eliminate isolated unreliable or noisy motions. To characterize the long-term motion feature, the Motion History Image (MHI) is employed as basic representations of the interest point. This new feature is named as Hierarchical Filtered Motion Field (HFMF) and works well in crowd scenes. We believe the HFMF describes complementary aspects of video actions and this work will combine HFMF with the existing features of [6] [12] for action detection tasks.

Compared with classification task, action detection is more challenging. There are only a few works devoted to action detection task [9], [7], [24], [23], [4]. These works only use single type of features. Although multiple feature fusion was proved to be effective in action classification [13] [5], it is still an untouched problem to combine multiple features for action detection.

The difficulty of applying multiple features for action detection is two fold: First, existing fusion methods [13] [5] assumes that each sample has the same number of features. However, in action detection, different features correspond to different detectors, and the numbers of detected salient patches are usually different subject to different features. Second, detecting actions in the videos involves a searching process in x-y-t dimensions, which is very computationally expensive. Many existing feature fusion methods [5] are usually too slow for this task. This paper employs Gaussian Mixture Models (GMMs) to model heterogeneous features, and the probability of a given feature vector is estimated effectively based on the GMM model. To locate the action of interests, we employ branch-and-bound methods to find the optimal subvolumes which correspond the largest GMM scores. Note that although this paper only combines three features from [12], [19], [6], our method is a general framework and can be used to fuse more features [3], [17], [25].

3. ADAPTIVE ACTION DETECTION

Given a video sequence V, we employ different STIP detectors to detect a collection of local feature vectors $\{\mathbf{x}_p^m\}$. where $p \in V$ denotes the location of the feature, and m denotes the feature type with $1 \leq m \leq M$. We employ the Gaussian Mixture Model (GMM) to model the probability that $\mathbf{x}^m$ belongs to the given action. Suppose a GMM contains K components, the probability can be written as

$$Pr(\mathbf{x}^m|\theta^m) = \sum_{k=1}^{K} w^m(k)\mathcal{N}(\mathbf{x}^m; \mu^m(k), \Sigma^m(k))$$

where $\mathcal{N}(\cdot)$ denotes the normal distribution, and $\mu^m(k)$ and $\Sigma^m(k)$ denote the mean and variance of kth normal component for feature m. The set of all parameters of GMM model is denoted as $\Theta = [\theta^1, \theta^2, ..., \theta^M]$, where $\theta^m = \{w^m(k), \mu^m(k), \Sigma^m(k)\}$.

The advantages of GMM are that it is based on a well-understood statistical model, and it is easy to combine multiple features using GMMs. With GMM, we can estimate the probability that each feature vector $\mathbf{x}^m$ belongs to the background or the action of interest. Suppose there are C categories of actions with parameter of $\Theta_1, \Theta_2, \cdots, \Theta_C$. Each category corresponds to GMMs with M features $\Theta_c = [\theta_c^1, \cdots, \theta_c^M]$.

The parameters of GMM can be estimated using maximum likelihood estimation. A straightforward way is to independently train the model for each category and each feature. However, as shown by Reynolds [16], it is more effective to obtain $\theta_1^m, \theta_2^m, \cdots, \theta_C^m$ coherently by the use of a universal background model. Following [16] we first train a background model θ_0^m which is independent to all the vectors X^{all} using the mth feature. Then we adapt $\theta_1^m, \cdots, \theta_C^m$ from θ_0^m by EM algorithm in the following way.

We first estimate posterior probability of each $\mathbf{x}_i^m$ subject to the background model θ_0^m

$$p_k^c(\mathbf{x}_p^m) = \frac{w(k)\mathcal{N}(\mathbf{x}_p^m; \mu_0^m(k), \Sigma_0^m(k))}{\sum_j w(j)\mathcal{N}(\mathbf{x}_p^m; \mu_0^m(j), \Sigma_0^m(j)} \quad (1)$$

Then we can update $\mu_c^m(k)$ by

$$\mu_c^m(k) = \frac{1}{n_c} \sum_{\mathbf{x}_p^m \in X^c} p_k^c(\mathbf{x}_p^m)\mathbf{x}_p^m. \quad (2)$$

Although we can update Σ_c based on $p_k^c(\mathbf{x}_p^m)$, in practice we force $w_c^m(k) = w_0^m(k)$ and $\Sigma_c^m(k) = \Sigma_0^m(k)$, which is computationally robust.

The advantage of employing background model are twofold: First, adapting GMM parameters from background model is more computational efficient and robust. Second, updating based on background model leads to a good alignment of different action models over different components, which makes the recognition more accurate.

After obtaining the GMM parameters and a video clip V, we can estimate the action category by

$$c^* = \arg\max_c \sum_{m=1}^{M} \sum_{\mathbf{x}_p^m \in V} \log Pr(\mathbf{x}_p^m|\theta_c^m) \quad (3)$$

Next we discuss the action detection task. We use a 3D subvolume to represent a region in the 3D video space that contains an action instance. A 3D subvoume $Q = [x0, x1, y0, y1, t0, t1]$ is parameterized as a 3D cube with six degrees of freedom in (x, y, t) space. Spatial and temporal localization of an action in a video sequence is rendered as searching for the optimal subvolume. The spatial locations of the subvolume identify where the action happens, while the temporal locations of the subvolume denote when the action happens. Given a video sequence, the optimal spatial-temporal subvolume Q^* yields the maximum GMM scores

$$Q^* = \arg\max_{Q \subseteq V} \mathcal{L}(Q|\Theta_c) = \arg\max_{Q \subseteq V} \sum_m \sum_{p \in V} \log Pr(\mathbf{x}_p^m|\theta_c^m) \quad (4)$$

By assigning each patch a score $f(\mathbf{x}_p^m) = \log Pr(\mathbf{x}_p^m|\theta_c^m)$, Equation (4) can be solved by branch-and-bound algorithm [10] [24]. Branch-and-bound approach was first developed for integer programming problems. Lampert *et al.* [10] [2] showed that branch-and-bound can be used for object detection in 2D image base on an smart formulation. Yuan *et al.* [24] developped an efficient algorithm which generalizes branch-and-bound algorithm to 3D space of videos. In this paper, we perform max-subvolume search using the 3D branch-and-bound algorithm in [24], which is an extension of the 2D branch-and-bound technique [10]. The detailed technical description of the 3D branch-and-bound algorithm is omitted due to limited space.

4. EXPERIMENTAL RESULTS

To validate our action detection scheme, we collect a new dataset in Microsoft Research Redmond (we call it MSR-II dataset in this paper)[1], with cluttered background and multiple people in each frame. We do not use the CMU action dataset [9] since there is only a single sequence for training in it. Hu *et al.* [7] used videos from retailing surveillance, however, the dataset is confidential due to the privacy issue. Wang *et al.* [20] collected an dataset of social game events, but their problem is about classification but not detection. Our MSR-II dataset includes 54 video sequences, each of which contains several different actions, e.g., hand waving, clapping, and boxing. These videos are taken with the background of parties, outdoor traffic, and walking people. Actors are asked to walk into the scene, perform one of the three kinds of action, and then walk out of the scenes with these backgrounds.

[1] http://research.microsoft.com/∼zliu/ActionRecoRsrc.

Figure 1 shows the differences between KTH dataset (a) and MSR-II dataset (b). Note that in MSR-II dataset there are a lot of people in the scene and we need locate person with action of interest from the scene.

To evaluate the detection results of our model, we manually labeled the MSR-II dataset with bounding subvolumes and action types. By denoting the subvolumes ground truth as $\mathbf{Q}^g = \{Q_1^g, Q_2^g, \ldots, Q_m^g\}$, and the detected subvolumes as $\mathbf{Q}^d = \{Q_1^d, Q_2^d, \ldots, Q_n^d\}$, we use $HG(Q_i^g)$ to denote whether a groundtruth subvolume Q_i^g is detected, and $TD(Q_j^d)$ to denote whether a detected subvolume makes sense or not. $HG(Q_i^g)$ and $TD(Q_j^d)$ are judged by checking whether the overlapping is above a threshold ($1/4$ in our experiment).

$$HG(Q_i^g) = \begin{cases} 1, & \text{if } \exists Q_k^d, \quad s.t. \frac{|Q_k^d \cap Q_i^g|}{|Q_i^g|} > \delta_1 \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

$$TD(Q_j^d) = \begin{cases} 1, & \text{if } \exists Q_k^g, \quad s.t. \frac{|Q_k^g \cap Q_j^d|}{|Q_j^d|} > \delta_2 \\ 0, & \text{otherwise,} \end{cases}$$

where $|\cdot|$ denotes for the area of the subvolume , and δ_1, δ_2 are parameters to judge the overlapping ratio. In this paper, δ_1 and δ_2 are set as $1/4$.

Based on HG and TD, precision and recall are defined as

$$\text{Precision} = \frac{\sum_{i=1}^m HG(Q_i^g)}{m}, \text{Recall} \quad = \frac{\sum_{j=1}^n TD(Q_j^d)}{n}$$

Given a collection of detected subvolumes, we can compute the precision-recall values. By using different thresholds of the region scores $\sum_{\mathbf{x} \in Q} f(\mathbf{x})$, we apply the branch-and-bound algorithm multiple times and obtain the precision-recall curves for three actions in MSR-II dataset.

In MSR-II dataset, we use half of the videos for training and the remaining half videos for testing. We compare the detection results of each of the three features [12], [19], [6], and find that both Laptev's feature [12] and Hierarchical Filter Motion feature [19] can obtain reasonable detection results, while Dollar's feature [6] leads to bad detection results. The reason for the failure of Dollar's feature might be that the Gabor filter based features are heavily affected by the cluttered background, since most of the detected patches fall in the background instead of action of interests. Since Dollar's feature fails to detect some actions, we only compare results of two single feature detection and the multiple feature detection using our model. Figure 2 show the precision-recall curves for three features. It can be seen that hierarchical filter motion feature works better than Laptev's in handclapping and boxing, but slightly worse than Laptev's feature in handwaving. However, combining these two detectors, our multiple feature detection works significantly better than using any single features in all the three actions. It is also interesting to see that if we incorporate the inappropriate feature, the corresponding detection rate will decrease. The results confirm

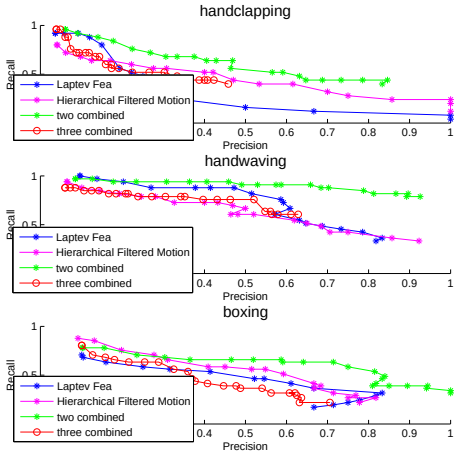

Fig. 2. Precision-Recall curves for MSR-II dataset.

Fig. 4. Our detector successfully detects the action even with heavy occlusion.

that *combining multiple relevant features will significantly improve the detection, while combining irrelevant feature might decrease the results.*

Figure 3 shows the action detection results using our multiple feature model. Even the background is cluttered and there are multiple persons in both close and distant view, our detector works well and can locate the action of interest very accurately. Moreover, our detector is robust subject to short-term occlusions. Figure 4 shows the detection results with heavy occlusion.

To compare our method with previous work, we test our algorithm on the public KTH dataset [18]. In KTH dataset, each video sequence exhibits one individual action from beginning to end, locating the actions of interest is trivial. In each video of the KTH dataset, we need not estimate Q since there is only one actor repeating the same action without background motions involved, and all the STIPs in the video are associated with the action. However, the classification task on KTH dataset can still show how our multiple feature fusion method outperforms single feature based methods. Following the standard experimental setting of KTH dataset as in [18], our method estimate the label of each video clip by (3). Table 1 shows that our feature fusion method outperforms

Table 1. Comparing the accuracy on KTH

Work	Accuracy
Schuldt *et al.* [18]	71.71%
Dollar *et al.* [6]	80.66%
Niebles and Fei-Fei [15]	83.92%
Huang *et al.* [8]	91.6%
Laptev *et al.* [12]	91.8%
Yuan *et al.* [24]	93.3%
Our work	**94.10%**

the single feature classification results.

5. CONCLUSION

This paper considers the problem of combining multiple features for action detection. We build a novel framework which combines GMM-based representation of STIPs and branch-and-bound based detection. We collect a new challenging dataset to validate our detection approach. The experimental results show that our approach can effectively detect the action even with cluttered background and partial occlusions. Our approach also outperforms the single feature classification results on KTH dataset.

6. REFERENCES

[1] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Conference on Computer Vision*, pages 1395–1402, 2005.

[2] M. Blaschko and C. Lampert. Learning to localize objects with structured output regression. In *European Conference on Computer Vision*, pages 2–15, 2008.

[3] O. Boiman and M. Irani. Detecting irregularities in images and in video. *IEEE International Conference on Computer Vision*, pages 462–469, 2005.

[4] L. Cao, Z. Liu, and T. S. Huang. Cross-dataset action detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2010.

[5] L. Cao, J. Luo, F. Liang, and T. Huang. Heterogeneous Feature Machines for Visual Recognition. *IEEE International Conference on Computer Vision*, 2009.

[6] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. *IEEE International Workshop on VS-PETS*, 2005.

[7] Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, and T. S. Huang. Action detection in complex scenes with spatial and temporal ambiguities. *IEEE International Conference on Computer Vision*, 2009.

[8] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. *IEEE International Conference on Computer Vision*, 2007.

[9] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. *IEEE International Conference on Computer Vision*, 2007.

[10] C. Lampert, M. Blaschko, and T. Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. In *CVPR*, 2008.

[11] I. Laptev and T. Lindeberg. Space-time interest points. *IEEE Conference on Computer Vision*, pages 432–439, 2003.

[12] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[13] J. Liu, S. Ali, and M. Shah. Recognizing human actions using multiple features. *IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.

[14] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". *IEEE Conf. on Computer Vision and Pattern Recognition*, 2009.

[15] J. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *British Machine Vision Conference*, 2006.

[16] D. Reynolds, T. Quatieri, and R. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1-3):19–41, 2000.

[17] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach: A spatio-temporal maximum average correlation height filter for action recognition. *IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.

[18] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. *ICPR*, 2004.

[19] Y. Tian, Z. Liu, L. Cao, and Z. Zhang. Hierarchical filtered motion field for action recognition in crowded videos. *Technical report, CCNY-ML-TR-001, The City College of New York*, 2009.

[20] P. Wang, G. D. Abowd, and J. M. Rehg. Quasi-periodic event analysis for social game retrieval. In *IEEE International Conference on Computer Vision*, 2009.

[21] S.-F. Wong, T.-K. Kim, and R. Cipolla. Learning motion categories using both semantic and structural information. *IEEE Conf. on Computer Vision and Pattern Recognition*, 2007.

[22] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. Rehg. A scalable approach to activity recognition based on object use. *IEEE International Conference on Computer Vision*, 2007.

[23] B. Yao and S. Zhu. Learning Deformable Action Templates from Cluttered Videos. *IEEE International Conference on Computer Vision*, 2009.

[24] J. Yuan, Z. Liu, and Y. Wu. Discriminative subvolume search for efficient action detection. *IEEE Conf. on Computer Vision and Pattern Recognition*, 2009.

[25] G. Zhu, M. Yang, K. Yu, W. Xu, and Y. Gong. Detecting video events based on action recognition in complex scenes using spatio-temporal descriptor. In *Proc. ACM international conference on Multimedia*, pages 165–174, 2009.

Fig. 3. Detection examples of MSR-II dataset. The bounding boxes denote the detected location using Branch-and-Bound search. The color of the bounding box denotes the action category: red for hand clapping, green for hand waving, and blue for boxing.