10-2009

Localizing volumetric motion for action recognition in realistic videos

Xiao WU

Chong-wah NGO
Singapore Management University, cwngo@smu.edu.sg

Jintao LI

Yongdong ZHANG

Localizing Volumetric Motion for Action Recognition in Realistic Videos

Xiao Wu[1,3], Chong-Wah Ngo[2], Jintao Li[1] and Yongdong Zhang[1]
[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[2]Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong
[3]Graduate University of Chinese Academy of Sciences, Beijing 100049, China
{wuxiao, jtli, zhyd}@ict.ac.cn, cwngo@cs.cityu.edu.hk

ABSTRACT

This paper presents a novel motion localization approach for recognizing actions and events in real videos. Examples include StandUp and Kiss in Hollywood movies. The challenge can be attributed to the large visual and motion variations imposed by realistic action poses. Previous works mainly focus on learning from descriptors of cuboids around space time interest points (STIP) to characterize actions. The size, shape and space-time position of cuboids are fixed without considering the underlying motion dynamics. This often results in large set of fragmentized cuboids which fail to capture long-term dynamic properties of realistic actions. This paper proposes the detection of spatio-temporal motion volumes (namely Volume of Interest, VOI) of scale and position adaptive to localize actions. First, motions are described as bags of point trajectories by tracking keypoints along the time dimension. VOIs are then adaptively extracted by clustering trajectory on the motion mainfold. The resulting VOIs, of varying scales and centering at arbitrary positions depending on motion dynamics, are eventually described by SIFT and 3D gradient features for action recognition. Comparing with fixed-size cuboids, VOI allows comprehensive modeling of long-term motion and shows better capability in capturing contextual information associated with motion dynamics. Experiments on a realistic Hollywood movie dataset show that the proposed approach can achieve 20% relative improvement compared to the state-of-the-art STIP based algorithm.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding - Video analysis

General Terms

Algorithm, Experimentation, Performance

Keywords

Human action recognition, Realistic videos, Motion subspace learning, Keypoint trajectory, Mean-shift clustering

MM'09, October 19–24, 2009, Beijing, China.

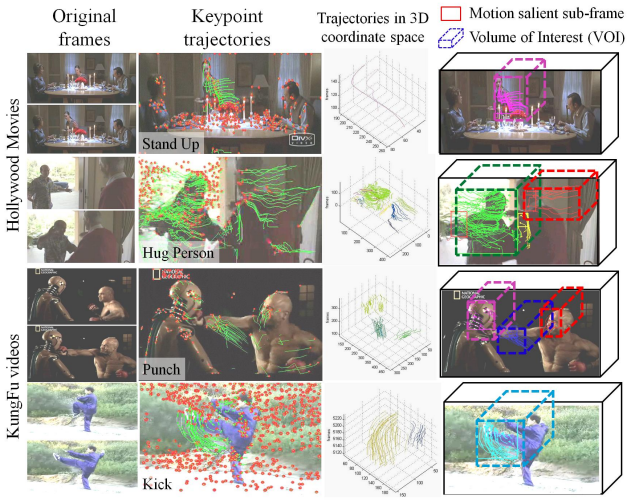

Figure 1: Examples of realistic human actions in Hollywood movies and KungFu videos. Actions in real videos exhibit large visual variation. This work proposes to use keypoint trajectories to track motion and detect 3D video volumes to localize actions.

1. INTRODUCTION

Automatically recognizing human actions in videos is increasingly receiving research attentions due to its great potentials for various industry applications, such as event-based video browsing, semantic indexing, video search and human-computer interaction. Most early works focus on detecting actions in domain-specified videos such as surveillance and sports videos, as well as simplified action corpus (e.g. Weizmann action dataset used in [6, 7]). Recently, a number of works [1, 2, 3, 5] aim to recognize actions and events from realistic videos, such as movies, news videos and user-generated web videos. The types of human actions to being detected include single human behaviors such as Stand Up and multibody interactions such as People Hugging. However, realistic action recognition is highly challenging due to the presence of large intra-class variation, unconstrained camera viewpoint and clutter background, as illustrated in Fig. 1. Due to these visual variations, determining *when* and *where* actions happen in a video become difficult. This in turn also affects the effectiveness of feature extraction and pattern learning for action recognition. In this paper, we address the problem of *when* by position-independent localization and *where* by scale-adaptive volumetric detection, through the analysis of long-term motion dynamics.

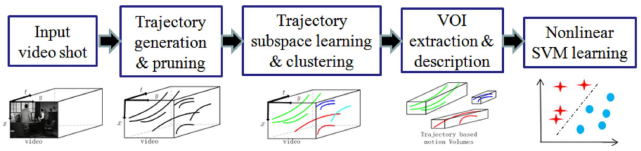

Figure 2: Flowchart of proposed algorithm for motion volume localization, modeling and learning.

To capture human actions, early works focus on extracting geometry shapes of human body from video [6, 7]. Since they rely on exact knowledge of the object contours or multiple view geometries, these schemes are not reliable for analyzing complicated real videos, in which geometry is hard to extract and prone to errors. Another line of research [3, 5, 10] employ space time interest points (STIP) based action modeling schemes. Feature points are located either by a 3D Harris corner detector [5] or Gabor filters [10]. The descriptors around those interest points are then computed and quantized into Bag of visual-Words (BoW) whose statistical distributions are used to represent the entire video sequence. Using the BoW, discriminative learning models (e.g. SVM) and generative models (e.g. pLSA) were adopted [3]. For example, in [5] actions are implicitly localized by small cubes and modeled using pixel statistics in fixed sized cubes. The problem with these methods is that fragmentized cubes cannot capture long-term dynamics of actions. More importantly, fixing the size of cubes is not adaptive to describe actions which appear at uncertain space-time location.

To adaptively localize spatiotemporal motion for precise action modeling, we propose trajectory generation to capture motion dynamics and subspace learning for action localization. Fig. 2 shows the overview of the proposed procedure. Bag of keypoint trajectories are first generated and then clustered into separate motion flows by motion subspace learning. Trajectories in the same cluster are expanded to form a 3D volume of adaptive size, namely **Volume of Interest (VOI)**. VOI descriptors are then extracted for action learning. Comparing with small and fragmentized STIP cuboids in [5], VOIs with larger spatiotemporal scope are more suitable for capturing macroscopic human behaviors, e.g. StandUp and GetOutCar. By utilizing fast trajectory computation, our approach is also more efficient than STIP based method. We validate our approach on realistic action videos (Hollywood movie shots from [5]), on which using complementary features the dynamic VOIs outperforms STIP based method [5] by a margin of 20%.

2. VOLUMETRIC MOTION LOCALIZATION

In this section we present a novel space-time motion localization algorithm by keypoint trajectories clustering. The detected video volumes are called Volumes of Interest (VOIs), from which features are extracted for video action modeling.

2.1 Trajectory Generation and Pruning

Trajectory Generation: We use keypoint trajectories to track all motions in video, including human body actions. Similar to recent work [11] which uses kinematic patterns of trajectories to detect copied videos, we employ pyramid Lucas-Kanade keypoint detection and tracking implementations in OpenCV [12]. As shown in Fig. 1, generated keypoint trajectories are capable of tightly tracking motions. To prevent from introducing large camera motion, we set the maximal trajectory length to 75 frames (equals to about

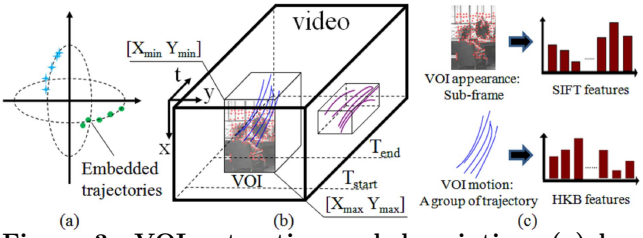

Figure 3: VOI extraction and description, (a) low dimensional embedding of trajectories, (b) determination of spatiotemporal boundary of 3D VOI by using the minimal and maximal coordinates of trajectories in a cluster, (c) describing VOI with SIFT and HKB (Histogram of Keypoint Behaviors).

3 seconds of duration). When trajectories are cut off at shot boundary or when reaching max length, a new session of keypoint detection and tracking starts to generate new trajectories. We define keypoint trajectories as below.

DEFINITION 2.1. *A keypoint trajectory T with length l is defined as a sequence of point coordinates, where t denotes its start frame in video.*

$$T_t^{t+l} = \overrightarrow{p_t p_{t+1} ... p_{t+i} ... p_{t+l}} = \{(x_t, y_t), (x_{t+1}, y_{t+1}), ..., (x_{t+i}, y_{t+i}), ..., (x_{t+l}, y_{t+l})\} \quad (1)$$

Trajectory Pruning: As observed in [2, 3] that actions and events tend to be dominated by significant motion dynamics, we conduct trajectory pruning to retain suspect action-related motions by filtering out motionless trajectories. A criterion $C(T)$ is defined for this purpose,

$$C(T) = [std(\overline{X}) + std(\overline{Y})]/2, \quad where \quad \overline{X} = \{x_1, x_2, ..., x_f\}, \quad \overline{Y} = \{y_1, y_2, ..., y_f\} \quad (2)$$

where $std()$ denotes the standard variance of coordinate sequence. Physical meaning of $std(\overline{X})$ and $std(\overline{Y})$ is the vertical and horizonal motion intensity of T, respectively. We threshold $C(T)$ to filter out nondistinctive trajectories. The purpose is to keep only those suspect actions of trajectories for further analysis and thus the threshold is set empirically. In the experiment, by setting the threshold $\Theta \leq 8.5$, in average 12.9% trajectories are removed.

2.2 Trajectory Clustering

We adopt motion subspace learning to project trajectories onto a low dimensional space, where trajectories with different behaviors are discriminatively distributed. This will facilitate trajectory clustering for localizing various motion patterns in a video shot [8].

Motion Matrix: First we construct a motion matrix $M_{2f \times p}$ using coordinates of all trajectories, where f is the number of frames and p denotes the number of trajectories.

$$M_{2f \times p} = [T^1_{2f \times 1}, T^2_{2f \times 1}, \cdots, T^i_{2f \times 1}, \cdots, T^p_{2f \times 1}], \quad i \in (1, p)$$
$$where \quad T^i_{2f \times 1} = [x^i_1, y^i_1, \cdots, x^i_j, y^i_j, \cdots, x^i_f, y^i_f]', \quad j \in (1, f) \quad (3)$$

Presenting trajectories in the form of motion matrix requires length normalization. To keep the originality of data, we adopt zero-padding (i.e. concatenating extra 0 to the end of trajectories) for length normalization as suggested in [4].

SVD decomposition is conducted on motion matrix $M_{2f \times p}$,

$$M_{2f \times p} = U_{2f \times K} S_{K \times K} V^T_{K \times p} \quad (4)$$

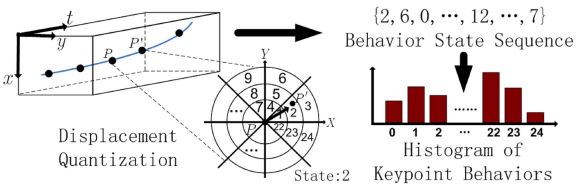

Figure 4: Quantifying a trajectory into Histogram of Keypoint Behaviors (HKB). The resulting 25D feature describes the statistical velocity and orientation of keypoints motion on a trajectory.

where we assume rank(M) is K. After normalizing each column of V, we use a unit vector $v_i (i = 1, ..., p)$ to represent the corresponding trajectory T_i in M. SVD decomposition is a transformation that projects a R^{2f} vector m_i (the i^{th} column of M) onto the R^K unit sphere which preserves the subspace property, as illustrated in Fig. 3(a). A subset of $m_i (i = 1, ..., p)$ spans a subspace of the same rank of the corresponding subset of $v_i (i = 1, ..., p)$ [8]. After normalizing data onto a low dimensional sphere, a group of neighboring unit vectors are corresponding to trajectories with proximity and similar kinematic behaviors. Intuitively these trajectories probably come from identical component of body.

Since the number of moving parts is unknown in real videos, we adopt mean-shift clustering [9] on $V^T_{K\times p}$ to adaptively group similar v_i, as in Fig. 3(a). The algorithm accepts the bandwidth (i.e. average range of a cluster) as the parameter, which is empirically fixed to 350, a relatively small value for clustering similar trajectories.

2.3 Localizing Action Motions

To localize motions of action in a video, we use a volumetric 3D cube to encapsulate all trajectories in a cluster. The 3D cube is named as Volume of Interest (VOI). As demonstrated in Fig. 3(b), the boundary of a VOI is determined using the maximal and minimal coordinates $\{x, y, t\}$ of m_i in the cluster. As examples shown in Fig. 1, the bounding box of VOIs explicitly localize actions on both spatial and temporal domain. Each VOI contains a distinct motion pattern. Features of all VOIs in a shot are extracted for learning.

3. FEATURE EXTRACTION AND ACTION CLASSIFIERS LEARNING

Localized motion volumes are described by the widely used volumetric features, e.g. 3D Histograms of Oriented Gradient (HoG) [5] and SIFT. A 3D HoG is used to describe a VOI. SIFT features are extracted from a sub-frame at the center of VOI (see Fig. 3(c)). In addition we describe trajectories as Histograms of Keypoint Behaviors (denoted as HKB), as shown in Fig. 4. HKB quantizes the displacements between any two adjacent keypoints of a trajectory into 25 states according to their velocities and orientations, as shown in Fig. 4. With HKB, each trajectory is represented as a histogram of 25 states characterizing its 3D motion in VOI. For efficient and compact representation, Bag of visual-Words model (BoW) is further applied to quantify the VOI features. Specifically, the HoG, SIFT, and HKB features of all VOIs in a video shost, respectively, are clustered to form three visual vocabularies, of each 1000 words, for describing the video shot as three BoW histograms.

To train action classifiers, we adopt nonlinear SVM learning using pre-computed kernel matrix $K_{N\times N}$ as in [5]. N is

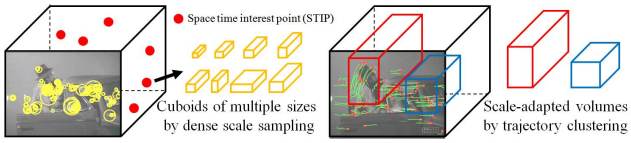

Figure 5: Comparison of STIP cuboids and VOIs in a video of GetOutCar. Multiple STIP cuboids of predefined scales are sampled to label motion (left), while VOIs localize actions of the person and car separately in two cuboids of adaptive scale (right).

the number of histograms for training. Using kernel matrix we are able to use single feature or combine multiple features mentioned in Sec. 3 to model video actions. An elementary kernel function (i.e. $K_{ij}, i, j = 1, ..., N$) is computed as,

$$K\left(H_i, H_j\right) = \exp\left[-\sum_{c\in\mathcal{C}} \frac{1}{A_c} D_c\left(H_i^c, H_j^c\right)\right] \quad (5)$$

where $H_i^c = \{h_{in}^c\}$ and $H_j^c = \left\{h_{jn}^c\right\}$ are two BoW histograms of visual words extracted in feature type c (e.g. SIFT or HKB descriptor) for the i-th and j-th samples respectively, whereas $D_c\left(H_i, H_j\right)$ is the χ^2 distance, namely

$$D_c\left(H_i^c, H_j^c\right) = \frac{1}{2}\sum_{n=1}^{N^c} \frac{\left(h_{in}^c - h_{jn}^c\right)^2}{h_{in}^c + h_{jn}^c} \quad (6)$$

and A_c is the average distance for normalization as in [5].

4. EXPERIMENTS AND DISCUSSIONS

The proposed approach is tested on the HOHA dataset [5], a benchmark human action dataset which contains 430 movie video shots (219 for training and 211 for test) belonging to eight action classes (see Fig. 6). We adopt one-against-all strategy to train 8 action classifiers and evaluate the performance using Average Precision (AP).

4.1 VOI versus STIP Cuboids

We first compare our proposed VOIs to the STIP cuboids based on [5]. Fig.5 shows an example for a video with the action 'get out of car'. Our approach requires neither scale selection [10] nor dense scale sampling [5] to determine the size of cubes to capture action. The merit of VOI lies at it encapsulates trajectories to roughly label the spatiotemporal range of an action. In HOHA dataset the average size of VOIs is $(100, 72, 60)$, which is much larger than size of most STIP cuboids (e.g. $(36, 36, 25)$ [5]). The average number of VOIs in a video shot is 49.5, which is much less than hundreds of STIP cuboids. It is worthnoting that [5] utilizes griding technique to segment a shot into grids from which features are extracted as channels. During their experiments, channels are exhaustively combined and evaluated. The best performance as a result of experimenting different ways of combination is presented in [5]. Using VOIs, such brute-force kinds of evaluation is not required since VOIs are scale and position adaptive to the underlying motion dynamics. For fair evaluation, we only compare our approach to the standard STIP in [5] without grid search.

In terms of speed efficiency, VOI detection is also more efficient than STIP detection. Using implementation of [5], it takes 132 seconds in average to detect space-time points from a video of 40 seconds duration. Using OpenCV based implementation, we detect VOIs in the shot within 41 seconds. This is because we use tracking to avoid keypoint detection on all frames which is highly time consuming.

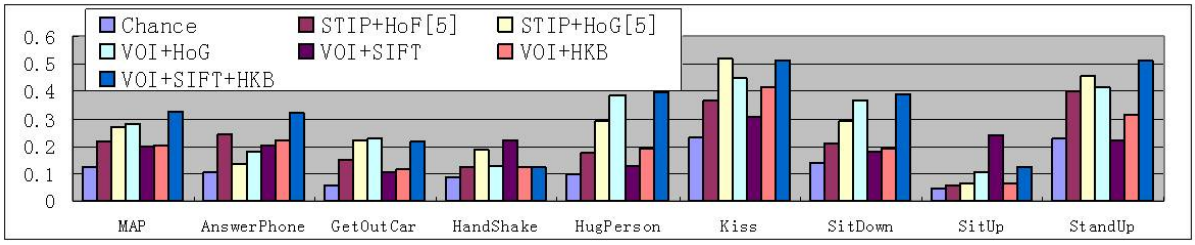

Figure 6: Comparison of APs between proposed method and [5] for recognizing action events in HOHA.

4.2 Action Recognition

We quantitatively compare the action recognition accuracy of VOI and STIP [5] with different BoW features. Fig.6 shows the performances for eight actions on HOHA. Basically, VOI shows consistently better performance than STIP for most actions when using same feature. VOI+HoG achieves mean AP (MAP) of 28.22%, while for STIP+HoG it is 27%. This indicates that action related features can be better extracted and modeled if the cuboids can cope with the underlying motion dynamics. STIP which often encodes an action with excessive discrete and fragmented cuboids, in contrast to VOI, fails in capturing motion in a holistic manner.

We also experiment the effectiveness of SIFT and HKB to describe VOI. As shown in Fig.6, separately employing SIFT or HKB indeed does not show satisfactory performance. By combining both features for VOI, which jointly takes into account the spatiotemporal dynamics and visual appearance, the best overall MAP performance is exhibited in the experiment. The BoWs of SIFT and HKB are combined in kernel matrix computation and the recognition achieves 32.45% in MAP. This shows 20% improvement compared to STIP+HoG (MAP=27%) of [5]. Among the eight actions in HOHA, VOI is particularly capable of capturing multiple body motion such as HugPerson, where their trajectories exhibit distinctive motion patterns to be recognized. However, for action such as HandShake, action related trajectories are almost motionless and as a result being filtered out during trajectory pruning stage. Recognizing this type of actions remain challenging by using either VOI or STIP.

In addition to HOHA dataset, we also apply our approach to KungFu videos crawled from the web. Fig.7 shows the detected VOIs on HOHA and Tai-Chi Kung-Fu videos. KungFu videos show various actions such as complex Tai-Chi martial arts which are difficult to extract and model with existing approaches. Our results using VOI show encouraging performance, where trajectories of human body motion are densely tracked and VOIs are able to encapsulate actions with scale and position adaptive volumes.

5. CONCLUSIONS AND FUTURE WORKS

We have presented a novel motion localization approach for realistic action recognition. By robust keypoint trajectory and motion subspace learning, our approach locates volumes of motions which can cope with the long-term motion dynamics of actions. We demonstrate that using scale and position adaptive motion cuboids, visual features are more holistically extracted and can lead to better recognition performance. Our future works include recognizing complex actions in martial art videos such as KungFu movies, which are difficult to index and tag even with manual labeling.

6. ACKNOWLEDGEMENTS

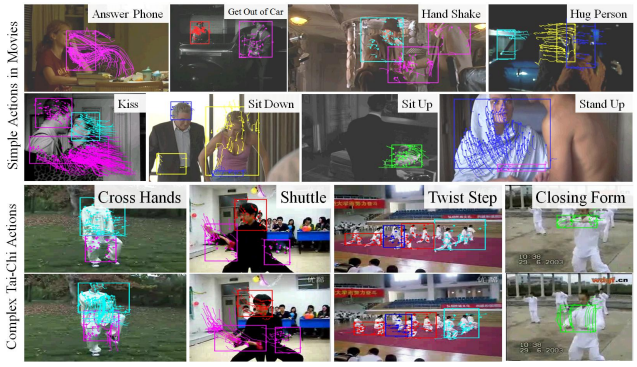

Figure 7: Detected VOIs in realistic videos of HOHA movies [5] and Tai-Chi KungFu videos on the web. Trajectories can capture action motion and VOIs localize moving human bodies. Note that action TwistStep of the crowd is not correctly captured because human bodies are in very small scale.

This work was supported by the National Basic Research Program of China (973 Program, 2007CB311100), National High Technology and Research Development Program of China (863 Program, 2007AA01Z416), National Nature Science Foundation of China (60873165 & 60802028), Beijing New Star Project on Science & Technology (2007B071) and the Co-building Program of Beijing Municipal Education Commission.

7. REFERENCES

[1] J. Sun, X. Wu, SC. Yan, LF. Cheong, TS. Chua and J. Li. Hierarchical spatio-temporal context modeling for action recognition. *CVPR*, 2009.
[2] F. Wang, Y. Jiang, and C. Ngo. Video event detection using motion relativity and visual relatedness. *ACM Multimedia*, 2008.
[3] J. Liu, J. Luo, et al. Recognizing realistic actions from videos 'in the Wild'. *CVPR*, 2009.
[4] B. Morris, et al. A survey of vision-based trajectory learning and analysis for surveillance. *TCSVT*, 2008.
[5] I. Laptev, M. Marszałek, C. Schmid, et al. Learning realistic human actions from movies. *CVPR*, 2008.
[6] D. Batra, T. Chen and R. Sukthankar. Space-Time shapelets for action recognition. *IEEE WMVC*, 2008.
[7] L. Gorelick, M. Blank, E. Shechtman, et al. Actions as space-time shapes. *TPAMI*, 2007.
[8] R. Tron, et al. A benchmark for the comparison of 3D motion segmentation algorithms. *CVPR*, 2008.
[9] Y. Cheng, et al. Mean shift, mode seeking, and clustering. *TPAMI*, 1995.
[10] P. Dollar, V. Rabaud, et al. Behavior recognition via sparse spatio-temporal features. *In VS-PETS*, 2005.
[11] X. Wu, Y. Zhang, Y. Wu, J. Guo and J. Li. Invariant visual patterns for video copy detection. *ICPR*, 2008.
[12] OpenCV: sourceforge.net/projects/opencvlibrary.