Copyright © 2012 KSII

Spatio-temporal Semantic Features for Human Action Recognition

Jia Liu[1,2], Xiaonian Wang[1], Tianyu Li[1] and Jie Yang[1]
[1] Institute of Image Processing and Pattern Recognition,
Shanghai Jiao Tong University, China
[e-mail: liujia1982@sjtu.edu.cn]
[2] Network and Information Security Key Laboratory,
Engineering College of the Armed Police Forces,Xi'an 710086, China
*Corresponding author: Jia Liu

Received July 4, 2012; revised August 27, 2012; accepted September 19, 2012; published October 29, 2012

Abstract

Most approaches to human action recognition is limited due to the use of simple action datasets under controlled environments or focus on excessively localized features without sufficiently exploring the spatio-temporal information. This paper proposed a framework for recognizing realistic human actions. Specifically, a new action representation is proposed based on computing a rich set of descriptors from keypoint trajectories. To obtain efficient and compact representations for actions, we develop a feature fusion method to combine spatial-temporal local motion descriptors by the movement of the camera which is detected by the distribution of spatio-temporal interest points in the clips. A new topic model called Markov Semantic Model is proposed for semantic feature selection which relies on the different kinds of dependencies between words produced by “syntactic ” and “semantic” constraints. The informative features are selected collaboratively based on the different types of dependencies between words produced by short range and long range constraints. Building on the nonlinear SVMs, we validate this proposed hierarchical framework on several realistic action datasets.

Keywords: action recognition, spatio-temporal features, topic model, markov model

http://dx.doi.org/10.3837/tiis.2012.10.011

1. Introduction

Considerable progress has been made in classification of action is receiving more and more attention in computer vision community. Many existing methods [1][2][3] obtain high classification score for simple action sequences with exaggerated motion, static and uniform background in controlled environment.

Recent interest in human action recognition research has been shifted to more realistic action databases, such as sports broadcasting videos [4], home videos on YouTube [5] and film videos [6]. All difficulties associated with object detection and classification task, such as large intra-class variations, poor lighting, partial occlusions and cluttered background, may also be encountered in action recognition problem. Therefore, the problem of recognizing actions in these videos is challenging.

In this paper, we propose a novel method to address these problems partly for recognizing actions in an unconstrained environment. Firstly, we present a novel trajectory representation which is extract by the Affine SIFT points. Based on these trajectories, trajectories descriptors are computed to retain local motion information, trajectory shape information and appearance information. We also consider local features extracted based on spatio-temporal interest points as they play a complementary rule in human action features. One of our contribution is that we develop a feature fusion method to combine spatial-temporal local motion descriptors by the movement of the camera which is detected by the distribution of spatio-temporal interest points in the clips. To select more informative and discriminative features for action recognition, a new topic model called Markov semantic model (MSM) is proposed for feature selection. The informative features are selected collaboratively in a high dimensional feature space.

Extensive experiments were conducted to evaluate the effectiveness of the proposed framework using realistic action datasets including the KTH dataset [10], the YouTube Dataset [5], the UCF Sports dataset [4] and the HOHA2 dataset [6]. Our results demonstrate that the proposed methods achieving comparable results.

2. Related Work

Local spatio-temporal interest points features. Recent work in activity recognition has been largely based on local spatio-temporal features. Many of these features seem to be inspired by the success of statistical models of local features in object recognition. Local features are first detected by some interest point detector running over all locations at multiple scales. Local maxima of the detector are taken to be the center of a local spatial or spatio-temporal patch, which is extracted and summarized by some descriptors. Most of the time, these features are then clustered and assigned to words in a codebook, allowing the use of bag-of-words models from statistical natural language processing. Laptev and Lindeberg [11] propose a space-time interest point operator that detects local structures in space-time that image observations have large local variations in both space and time. Schüldt et al. [10] train an SVM classifier based on these space-time features for recognizing human actions. Dollár et al. [12] propose space-time interest point detector based on a set of linear filters, and use these local features with k-nearest neighbor classifier for action recognition. Wong and Cipolla [13] utilize the global information to yield a sparser set of interest points for motion recognition. Willems et al. [14] present the spatio-temporal interest points that are at the same time scale-invariant (both

spatially and temporally). Liu and Shah [15] exploit mutual information maximization techniques to learn a compact set of visual words. Niebles and Li [16] combine shape information with local appearance features by building a hierarchical model that can be characterized as a constellation of bag-of-features. Recently, a compact local descriptor of video dynamics proposed by Derpanis [17] in the context of action spotting is introduced based on visual space-time oriented energy measurements. However, features from interest points are limited in temporal scalability, therefore are inadequate for describing longer-term movements. Alternatively, long-term motion can be described by the trajectories through tracking key points. Here the short (simple) motion means the action primitive that is an atomic movement which can be described at the limb level. The long term motion consists of action primitives (short movements) and describes a, possibly cyclic, whole-body movement.

Spatio-temporal trajectory-based features. Trajectory-based action recognition has been extensively studied in the past few years. These proposed algorithms typically differ on how to encode the dynamics of trajectories for subsequent processing. Much of the traditional trajectory tracking work has often been based on object centroid or bounding box trajectories. In a more sophisticated extension of such approaches, Jiang and Martin [18] build a graph template of keypoint motion, and match it to regions. While this technique includes keypoints with potentially significant spatial and temporal extent, is expensive to run exactly. Ross et al.[8] analyzing feature trajectories using dense clouds of KLT [19] feature tracks to build a velocity history feature for representation. Sun et al. [17] used similar techniques to model SIFT-feature trajectories. However, they find a fixed-dimensional velocity description using the stationary distribution of a Markov chain velocity model. The stationary distribution is closer to a velocity histogram. Bregonzio ect. [9] compute trajectories of key-points using two techniques: the Pyramid Kanade-Lucas-Tomasi (KLT) tracker [19] and the SIFT matching [7]. In this paper, we use Affine-SIFT (ASIFT) detector and (KLT) tracker for trajectory extraction. We propose to capture visual motion patterns by extracting the trajectories of the Affine-SIFT [20] salient points and optical flow, and then model the spatio-temporal information using several descriptors based on these trajectories.

Latent topic models for action recognition. As one of the generic models, topic model [21] has been successfully used to discover object categories without prior segmentation. Recently, successes have been made in adopting generative topic models with "bag-of-words" framework in solving various recognition problems in computer vision. Niebles et al.[22] use latent dirichlet allocation (LDA) and probabilistic latent semantic analysis (pLSA) for human action categories and location. Wong et al.[23] extends pLSA to capture both semantic (content of parts) and structural (connection between parts) information for motion category recognition. Zhang et al. [24] proposed a new approach structural pLSA (SpLSA) to model explicitly word orders by introducing latent variables for human action categorization. Wang et al. [25] presented two semi-latent hierarchical topic models such as S-LDA and S-CTM for action recognition based on motion words. Hospedales et al. [26] proposed a Markov Clustering Topic Model (MCTM) which builds on existing Dynamic Bayesian Network models (e.g. Hidden Markov Models (HMMs)) and Bayesian topic models (e.g. Latent Dirichlet Allocation) for mining behavior in video. Most of above method using Topic model for classification, the latent topics in their models directly correspond to class labels. Actually, topic model such as LDA [27] can be also used for feature selection. Multi-Class Delta Dirichlet Allocation (MC-DLDA) topic model which was proposed in [9] for feature selection. However, they did not consider the relationship between visual words. In this paper，we proposed a new topic model called Markov Semantic Model for feature selection. Different

kinds of dependencies between words are explored by "syntactic" and "semantic" constraints. This method retains correlation among features and selects them collaboratively.

3. Action Representation

We use two types of features for action representation. Fisrt we extract the trajectories of the salient points to capture visual motion patterns and model the long-duration motion characteristics residing with these trajectories. Secondly, we also consider spatio-temporal interest point based descriptors which capture short (simple) movements within a short temporal for complement. The framework of action representation is shown in **Fig. 1**.

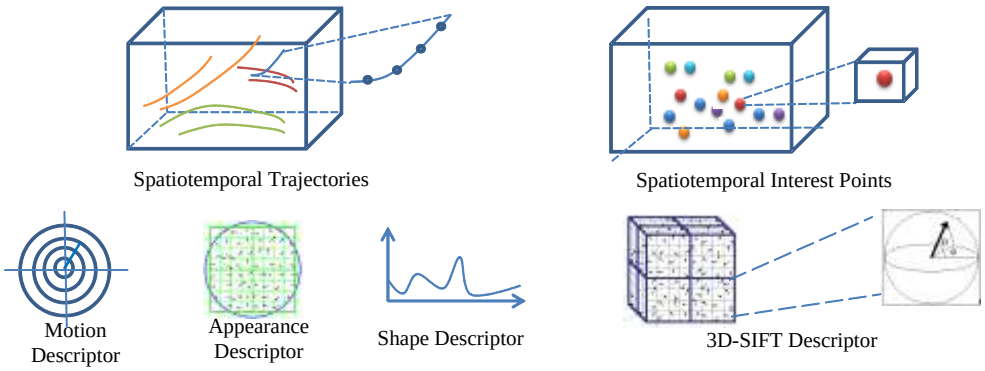


Fig. 1. Schematic diagram on action representation.

Both trajectories features and points features are considered in our framework. For trajectories, the three descriptors residing with long-term trajectories are 1) the point-level context (SIFT average descriptor), 2) trajectory motion context (orientation and magnitude), and 3) shape context (Fourier descriptors). For the spatio-temporal interest points, 3D SIFT descriptors was used for short-term representation. Note that our action trajectories features method is similar to the Ref.[9][17][40]. In their work, the SIFT descriptor and motion context information has been proven successfully in human action representation. Therefore, in this paper, following the framework in these work, we give the detail of the representation method in below sections.

3.1 Spatio-temporal Trajectory Features

Trajectory extraction – In realistic video, reliable spatially salient point detection and tracking algorithms are very critical for the modeling of trajectory patterns. We adopt the ASIFT [20] detector for salient point detection. The fully affine-invariance and scale-invariance properties of ASIFT render it a better choice as compared to other techniques such as the SIFT detector [7] and Harris detector. Because Kanade-Lucas-Tomasi (KLT) feature trackers [8] can provide dense trajectories which are complimentary information to ASIFT trajectories. These two trackers are applied independently to a video clip so that we can obtain trajectories as dense as possible even in low textured videos. For simplicity, we mainly describe the Affine SIFT trajectories extraction method, the Kanade-Lucas-Tomasi (KLT) feature trajectories can be found in [8]. The detail that ASIFT is proven fully affine invariant will be found in [20].

The fully affine invariant will produce more rich and reliable trajectories in realistic videos. The trajectory extraction process is based on the pairwise ASIFT matching over consecutive frames. For the frames $\{f_1,\ldots,f_k\}$ of a video sequence denoted V with k frames, we establish all the ASIFT point matches between f_i and f_i+1, for $0< i < k$+1. Matches that extend over several frames then form a motion trajectory of the ASIFT salient point. **Fig. 2** (a) shows the motion trajectories for a "running" action example from the KTH Dataset. The yellow points denote the matched ASIFT salient point at the current frame and the green curves denote the trajectories of the matched ASIFT salient points.

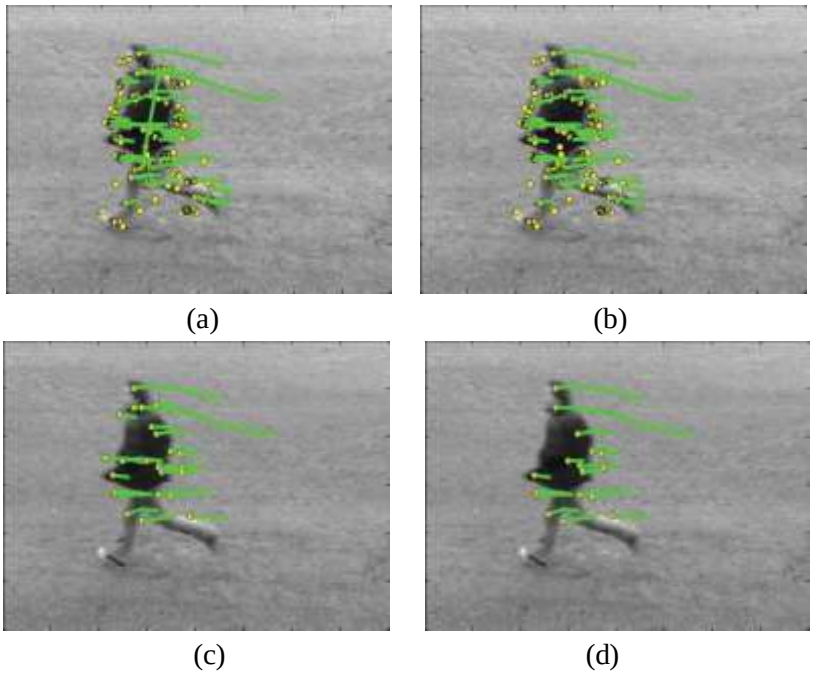


Fig. 2. Trajectories pruning procedure for one example clip. (a) All the detected trajectories. (b) Result of the first pruning step. (c) Result of the second pruning step. (d) Result of the final pruning step.

Trajectory pruning - Not all the extracted trajectories in a video are useful for action representation. For example, in **Fig. 2** (a), some of trajectories which are detected by the noise match points did not represent the action movement need to be removed in order to retain the most relevant trajectories for describing the actions of the human body.We consider a three-step trajectory pruning process.

1) For any ASIFT salient point p in frame i, there can be maximally one candidate match point p' in frame i+1, and p' must be located within a $N \times N$ (we set N = 16 in all the experiments) spatial window around point p. This windowing approach ensures that the trajectory may automatically end when reaching the shot boundaries or with considerable occlusions. **Fig. 2** (b) shows the result of the first pruning step.

2) To further remove possible noisy trajectories and reduce the chance of long trajectories mixing up with successive motions, we restrict the length L of any valid trajectory to be $L_{min} < L < L_{max}$. In this work, we set L_{min} = 5 and L_{max} = 25, which correspond to 0.2~ 1 second in duration. **Fig. 2** (c) shows the result of this pruning step. The length threshold of valid

trajectory ensures noisy trajectories that are consist of only several successive match points were detected. As can be seen in the Fig. 2(c), most of the noisy trajectories were removed from the Fig. 2(b). It is worth noting that L_{min} and L_{max} are fixed in our work. Although this is certainly valid, it may not be the optimal choice. For instance, different types of motion such as "running" and "jumping", the valid length should be different. The valid length is also affected by the rate at which the action is recorded. A more general approach would be to determine the valid length from the type of motion and rates of execution. Determining these values is however non-trivial which is the reason why we have restricted ourselves to the fixed values.

3) To remove the trajectories from the background, for each trajectory consists of L key points $\{(x_1, y_2), (x_1, y_2), \ldots, (x_L, y_L)\}$, we define an average trajectory path length with framewise displacement:

$$T_{avg\_path} = \frac{\sum_{j=1}^{L-1}\left((x_{j+1}-x_j)^2-(y_{j+1}-y_j)^2\right)^{1/2}}{L-1} \tag{1}$$

If $T_{avg\_path} < T_{path}$,(where $T_{path} = 0.5$ pixel for our work), the trajectory will be removed. The reason of the final step is that we assume the valid trajectory should have a long enough average displacement which can describe a long-term movement. An example of remaining trajectories after the last pruning step is shown in **Fig. 2**(d). It shows clearly that the region of interest corresponds accurately to where the action takes place.

Trajectory Shape – In order to give the shape signatures of action trajectories, we use a Fourier descriptor [38] to describe these trajectories. Given L key points $\{(x_1, y_1), (x_2, y_2), \ldots, (x_L, y_L)\}$ along a trajectory, we first give the centroid distance function which is expressed by the distance of the boundary points from the centroid (x_c, y_c) of the trajectory:

$$r_i = \sqrt{([x_i - x_c]^2 + [y_i - y_c]^2)}, \quad i = 1, 2, \ldots, L \tag{2}$$

In which x_c, y_c are averages value of x and y coordinates respectively. All of the movement trajectories are scale to the same number of points. In this paper, we assuming that there exist N points in the normalization stage, then the Fourier transform of r_i, $i = 0, 1, \ldots, N\text{-}1$ can be given by:

$$u_n = \frac{1}{N}\sum_{i=0}^{N-1} r_i \exp\left(\frac{-j2\pi\, n\, i}{N}\right), \quad n = 0, 1, \ldots, N-1 \tag{3}$$

The coefficients u_n are called Fourier descriptors (*FD*) of the trajectory which denoted as FD_n, $n = 0, 1, \ldots, N\text{-}1$. In order to achieve invariance for rotation, we ignored the phase information and only the magnitudes $|FD_n|$ are used. Scale invariance can be achieved by dividing the magnitudes by the Direct Current (DC) component, i.e., $|FD_0|$. Then the Fourier descriptors of a trajectory is given as follow:

$$\mathbf{f} = \left[\frac{|FD_1|}{|FD_0|}, \frac{|FD_2|}{|FD_0|}, \ldots, \frac{|FD_{N-1}|}{|FD_0|}\right] \tag{4}$$

Note that, comparison with the shape descriptor used in Ref.[9], this descriptor is a global descriptor and each trajectory is represented by an N-1 dimension vector.

Trajectory Motion – In order to capture the displacement of movement, we compute a series of displacement vectors $\boldsymbol{d} = \{d_1, d_2, \ldots, d_{L-1}\}$ for a single trajectory t of length L, where the d_i is the displacement vector between the two consecutive matching Affine-SIFT interest points:

$P = (x_l, y_l)$, $P_0=(x_{l+1}; y_{l+1})$ along a trajectory. We consider both the magnitude and orientation information to obtain a reasonable descriptor for dispalcement vectors. For magnitude, the displacement vector $\boldsymbol{d}$ is first normalized by the largest displacement magnitude $\boldsymbol{d}_{max}$ residing with the same trajectory, we set 4 uniform quantization levels. For orientation, the full circle are divided into 12 equal sectors, each subtending 30^o. Finnaly, we get the combination of magnitude and orientation quantization results in 48 bins in polar coordinate. The formulated track descriptor is both scale-invariant and direction-invariant. Similar motion descriptor is also used in Ref.[9] and Ref [17]. It is worth noting that our motion descriptors have more bins which make the descriptors more powerful than theirs.

Trajectory Appearance. The appearance context information of the trajectories is measured as the average of all the SIFT descriptors along the extracted trajectory. For a motion trajectory of length L, the SIFT average descriptor S is related to all SIFT descriptors $\{S_1, \ldots, S_L\}$ along this trajectory :

$$\boldsymbol{S} = \frac{\mathbf{1}}{\boldsymbol{L}} \sum_{i=1}^{L} \boldsymbol{S}_i \tag{5}$$

In Ref. [9] and Ref. [17], the appearance descriptor is used for trajectory representation. The point-level context not only ensures that the local image patches residing on the trajectory are stable, but also offers a robust representation for certain aspect of visual content within the video. Moreover, the SIFT descriptors along the trajectory is extract frame by frame, the average descriptor also encode the temporal information in some aspect.

Bag of Words (BOW) Representation. For each trajectory in video, three types of descriptors S (Shape) , M (Motion) and A (Appearance) are normalized and concatenated to form a global descriptor $G = [S, M, A]$. BOW method is employed for fusing these descriptors. We quantize the global descriptors G for all trajectories using K-means to obtain a codebook with 1000 words and assign each trajectory a codeword.

3.2 Spatiotemproal Interest Points Features

Following the work of [9] , the spatiio-temporal interest point features contain complementary information to trajectory descriptors, we also consider local features extracted based on interest points for action representation. First, interest points are detected using the method of [1] which tends to correspond to the main contributing body parts to the action being performed. Moreover, method in Ref [1] often produces more spatiotemporal interest points than Ref.[12]. 3D-SIFT descriptor [28] are extracted around the interest points at which the local maximal of detector response. The 3D SIFT descriptor is able to better represent the 3D nature of video data in the application of action recognition. Similar to trajectory representation, we also build a codebook V_p with 1000 words by performing K-means to a subset local features from the training data. The Bag-of-words framework is also used to describe each video clip.

3.3 Feature Fusion

As interest points features contain complementary information to trajectories features. We fuse trajectory based descriptors with 3D interest point based descriptors according the presence of camera movement. Since the spatio-temporal interest points can capture short movements within a short temporal window, the presence of moving camera is detected by the distribution of spatio-temporal interest points in the clips. Given a video clip with spatio-temporal interests points extracted by [1], we first divide the whole video clip with dimensions X, Y, T into S subcuboids, for each subcuboid, the density of the points in the

subcuboid is given by $D_i^{sub} = N_i^{sub} / V_{sub}$, i=1,…,S, where N_i^{sub} is the number of the points in the subcubiods i, V_{sub} is the volume of the subcuboid with spatial and temporal extents σ, σ, τ. The moving detection of the clip is given by,

$$M = \begin{cases} 1, & \sum_{i=1}^{S} f_i > S/2 \\ 0, & otherwise \end{cases}, \quad \text{where} \quad f_i = \begin{cases} 1, & D_i^{sub} > T \\ 0, & otherwise \end{cases} \tag{6}$$

fi is the binary feature depends on the threshold T of the density. We set $\sigma = X/3, \tau = T/3, S = 3 \times 3 \times 3 = 27, and\ T = \sum_{i=1}^{S} N_i^{sub} / XYT$.

If the majority of the frames contain global motion (M=1), we regard the clip as being recorded by a moving camera. If camera motion can be detected, interest point based descriptors are less meaningful so only trajectory descriptors are employed, resulting the final codebook $V = V_t$ with 1000 visual words. For clips without camera movement (M=0), both interest point and trajectory based descriptors can be computed reliably and thus both types of descriptors are used for recognition, resulting in a final codebook V = [V_t, V_p] with 2000 visual words. We also notice that Ref.[39] introduces a novel descriptor based on motion boundary histograms, which is robust to camera motion.

4. Feature Selection

To select more informative and discriminative features for action recognition, we propose a Markov Semantic Models (MSM) for feature selection. This model can be viewed as the extension of MC-LDA[9]. Although MC-LDA has achieved promising results in the application of feature selection for action recognition [9], the original model ignores the relationship between words.

Our model based on the assumption that not all the visual words provide semantic content for action due to the noises caused by camera movements or changing view-point and illumination, and low resolution. Visual words that play different roles are treated differently in video processing. Our approach relies on the different kinds of dependencies between words produced by “syntactic” and “semantic” constraints. “Syntactic” constraints introduced by noises result in relatively short-range dependencies, spanning several words but not going beyond the limits of a small set of visual words. “Semantic” constraints result in long-range dependencies: different words within a video are likely to have similar content with similar words.

4.1 Markov Semantic Model

We propose a new composite model in which the syntactic component is an hidden markov model (HMM) and the semantic component is a topic model. The model is defined in terms of three sets of variables: a sequence of words $\mathbf{w} = \{w_1, \ldots, w_n\}$, with each w_i being one of N_W words, a sequence of topic assignments $\mathbf{z} = \{z_1, \ldots z_n\}$, with each z_i being one of N_T topics, and a sequence of classes $c = \{c_1, \ldots, c_n\}$, with each c_i being one of C feature classes. One class, say $c_i = 1$, is designated the “semantic” feature class. The z_{th} topic is associated with a distribution over words $\varphi^{(z)}$, each class $c \neq 1$ is associated with a distribution over words $\varphi^{(c)}$, each document d has a distribution over topics $\theta^{(d)}$, and transitions between classes c_{i-1} and c_i follow a distribution $\pi^{(c_{i-1})}$. However, different from existing topic models such as LDA[27] and HMM-LDA[29] which assume uniform proportion of topic mixture for each video clip,

the topic model part of the MSM also aims to constrain topic proportion non-uniformly and on a per-clip basis. More precisely, for each video clip belonging to action category A_a, we model it as a mixture of: (1) N_s topics which are shared by all A_a category of actions, and (2) N_a topics which are uniquely associated with action category A_a. (3) N_c topics which are associated with feature class C. Therefore, the total number of topics will be $N_T = N_s + \sum_{a=1}^{Aa} N_a + N_c$. The structure of the proposed MSM is shown in **Fig. 3**.

In MSM, the non-uniform proportion of topic mixture for a single clip **w** is enforced by its action class label a and the hyperparameter α^a for the corresponding action class a. Generative process of sampling video clips is given as follows:

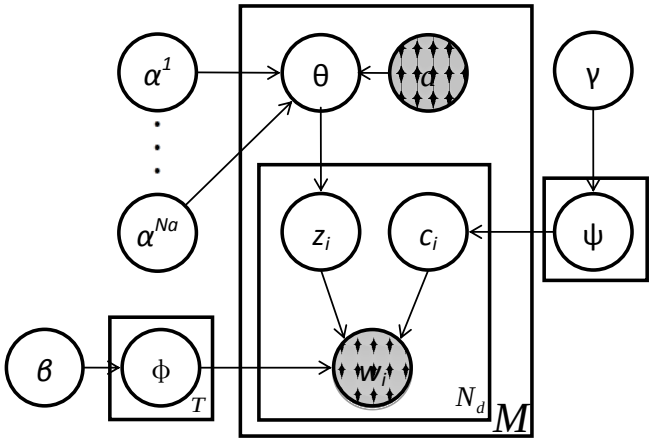


Fig. 3. Graphical model of Markov Semantic Model.

1. Draw a Dirichlet word-topic distribution $\phi \sim Dir(\beta)$ for every topic.
   Draw the rows of transition matrix $\psi_c \sim Dir(\gamma)$.
2. For each document $d = 1,\ldots, D$:
   (a) Draw a action class label $a_d \sim Multi(\varepsilon)$.
   (b) Given label a_d, draw a constrained topic distribution $\theta_{a_d} \sim Dir(\alpha^{a_d})$.
3. For each word in document d:
   (a) Draw a feature label $c_i \sim Multi\ (\psi_{c_{i-1}})$.
   (b) Draw a topic $z_i \sim Multi(\theta_{a_d})$.
   (c) If $c_i = 1$, draw $w_i \sim Multi(\phi^{(z)})$, else draw $w_i \sim Multi(\phi^{(c)})$.

Given the structure of the MSM and observable variables (clips $\mathbf{w}_d$ and action labels a_d), the full joint probability of a video (document) d in MSM is:

$$P(\mathbf{w}, \mathbf{z}, \mathbf{c}, \theta, \Phi, \Psi, a \mid \alpha, \beta, \gamma) = \prod_i P(w_i \mid z_i, c_i, \Phi) P(\Phi \mid \beta) P(z_i \mid \theta) P(\theta \mid \alpha, a) P(c_i \mid \Psi) P(\Psi \mid \gamma) \tag{7}$$

Although it is computationally intractable to perform inference and parameter estimation for the hierarchical Bayesian models, several approximation algorithms have been

investigated, e.g. Markov chain Monte Carlo(MCMC)[30], variational Bayesian inference[27] and expectation propagation [31]. In this paper, we use a Markov chain Monte Carlo algorithm to address the inference problem.

We need to derive the $P(z_i \mid \mathbf{w}, \mathbf{z}_{-i}, \mathbf{c}, a)$, the conditional distribution of a topic for the word $w_i = w$ given all other words' topic assignments, $\mathbf{z}_{-i}$ and $\mathbf{x}_{-i}$, to carry out the Gibbs sampling procedure for MSM.

$$\begin{aligned} P(z_i \mid \mathbf{w}, \mathbf{z}_{-i}, \mathbf{c}, a) &= P(z_i \mid \mathbf{z}_{-i}, \mathbf{w}_{-i}, w_i, \mathbf{c}, a) \\ &= \frac{P(z_i \mid \mathbf{z}_{-i}) P(w_i \mid \mathbf{z}, \mathbf{c}, \mathbf{w}_{-i}, a)}{P(w_i)} \\ &\propto P(z_i \mid \mathbf{z}_{-i}) P(w_i \mid \mathbf{z}, \mathbf{c}, \mathbf{w}_{-i}, a) \\ &\propto \begin{cases} \dfrac{n_{z_i}^{d_i} + \alpha^{a_i}}{\sum_T n_{z_i}^{d_i} + \alpha^{a_i}} \dfrac{n_{w_i}^{z_i} + \beta}{\sum_V n_{w_i}^{z_i} + \beta} & c_i = 1 \\ \dfrac{n_{z_i}^{d_i} + \alpha^{a_i}}{\sum_T n_{z_i}^{d_i} + \alpha^{a_i}} & c_i \neq 1 \end{cases} \end{aligned} \tag{8}$$

where $n_{z_i}^{d_i}$ is the number of words in document d_i assigned to topic z_i, $n_{w_i}^{z_i}$ is the number of words assigned to topic z_i that are the same as w_i, and all counts include only words for which $c_i = 1$ and exclude case i. We have obtained these conditional distributions by using the conjugacy of the Dirichlet and multinomial distributions to integrate out the parameters θ, ϕ and ψ.

Similarly conditioned on the other variables, each c_i is drawn from:

$$\begin{aligned} P(c_i \mid \mathbf{w}, \mathbf{z}, \mathbf{c}_{-i}, a) &= P(c_i \mid \mathbf{c}_{-i}, \mathbf{z}, \mathbf{w}_{-i}, w_i, a) \\ &= \frac{P(c_i \mid \mathbf{c}_{-i}) P(w_i \mid \mathbf{z}, \mathbf{c}_{-i}, c_i, \mathbf{w}_{-i}, a)}{P(w_i)} \\ &\propto P(c_i \mid \mathbf{c}_{-i}) P(w_i \mid \mathbf{z}, \mathbf{c}, \mathbf{w}_{-i}, a) \\ &\propto \begin{cases} \dfrac{(n_{c_i}^{c_i-1} + \gamma)(n_{c_i+1}^{c_i} + I(c_{i-1} = c_i) \cdot I(c_i = c_{i+1}) + \gamma)}{I(c_{i-1} = ci) + \sum_C n_{c_i+1}^{c_i} + \gamma_{c_i}} \dfrac{n_{w_i}^{z_i,-i} + \beta}{\sum_V n_{w_i}^{z_i,-i} + \beta}, & c_i = 1 \\ \dfrac{(n_{c_i}^{c_i-1} + \gamma)(n_{c_i+1}^{c_i} + I(c_{i-1} = c_i) \cdot I(c_i = c_{i+1}) + \gamma)}{I(c_{i-1} = ci) + \sum_C n_{c_i+1}^{c_i} + \gamma_{c_i}} \dfrac{n_{w_i}^{c_i,-i} + \delta}{\sum_V n_{w_i}^{c_i,-i} + \delta}, & c_i \neq 1 \end{cases} \end{aligned} \tag{9}$$

where $n_{w_i}^{z_i}$ is as before, $n_{c_i}^{c_i-1}$ is the number of transitions from class c_{i-1} to class c_i, and all counts of transitions exclude transitions both to and from c_i. $I(\cdot)$ is an indicator function, taking the value 1 when its argument is true, and 0 otherwise.

In our model, we set the number of shared topic to $N_s = 10$, and assigned each action category a unique topic $N_a = 1$. Moreover, we set the number of class topics $N_c = 3$ ($c_i = 1$ for semantic class and $c_i \neq 1$ for syntactic class). In this paper we do not estimate the hyperparameters α^a, β and γ instead they are fixed at 0.1, 0.01 and 0.5 respectively in the experiment.

4.2 Feature Selection using MSM

As visual features extracted from action videos can be very noisy and not well structured, those topics can be easily corrupted by noise. Using MSM enables us to learn N_t topics to represent natural grouping of semantic shared by all classes of actions or uniquely associated with one particular action category. Therefore, we first select the semantic visual words then learn the discriminative features from the N_s topics shared by all actions. Given the N_s shared

topics which are represented as an $N_w \times N_s$ dimension matrix Φ^S, the feature selection can be summarized into three steps:

1) For each feature words w_i in a video clip, if the corresponding class label $c_i = 1$, the feature w_i is preserved, if the class label $c_i \neq 1$, the feature w_i is discarded, all of the preserved w_i consist of the semantic feature vector $\boldsymbol{w_s}$.

2) For each word item v_k, $k = 1,\ldots N_w$, compute its maximum probability across all topics according to $p(v_k) = \max(\Phi^S_{k,1:N_s})$, then rank the value of $p(v_k)$ in ascending order to obtain a vector of words index $r(V)$ in which higher ranked words correspond to more discriminative features.

3) Given the number of selected features S, for each feature in $\boldsymbol{w_s}$, if the $w_i = v_k$ belong to the selected words vector, the feature w_i is preserved, otherwise, feature w_i is discarded.

After the feature selection process, the final features not only capture the semantic content for actions, but also preserve more discriminative information. The number of features selected for final classification is fixed at 300 for convenience.

5. Experiments

5.1 Datasets and settings

We have extensively applied our proposed approach to four datasets: KTH, UCF sport and UCF YouTube and HOHA2 datasets . **KTH action Dataset** [10] containing six types of human actions (walking, jogging, running, boxing, hand waving and hand clapping) performed several times by 25 subjects in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. Although KTH dataset is still a dataset with a simple background in controlled environment, it is the most widely used dataset in action recognition. In order to make a fair comparison of other methods and ours, we also use KTH Dataset to validate our method. **UCF Sport Actions Dataset** [4] contains 10 different types of human actions in sport broadcasting videos which show a large intra-class variability. The videos have different frame rate and image size. They last in average 5 seconds. **The Hollywood2 human action dataset** (HOHA2) [6] contains twelve actions (answer phone, drive car, eat, fight person, get out of car, handshake, hug, kiss, run, sit down, sit up and stand up), extracted from movies and performed by a variety of actors. There is a huge variety of performance of the actions, both spatially and temporally. **YouTube Dataset** [5] is the most extensive realistic action dataset available to public. It contains 11 action categories.For each category, the videos are grouped into 25 groups with more than 4 action clips in it. This dataset is very challenging due to large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions, etc. Clips have different frame rate by constant frame size of 320 by 240 pixels. The clips last in between 3 and 15 seconds. Examples of the four datasets are shown in **Fig. 4**.

In our experiment, we used a support vector machine (SVM) with Histogram Intersection kernel as a classifier [32]. Leave-One-Out Cross-Validation (LOOCV) was used for evaluate our algorithm. For KTH dataset, we followed the experimental setup of Schüldt et al. [10] with sequences divided into the training/validation setand the test set. For the HOHA2 movie actions datasets, we used the clean training dataset. The datasets were divided into 6 subsets, out of which 5 subsets were used for training and the remaining subset was used for testing. For the UCF Sport Actions datasets we followed the setting in [4][9]: one clip was used for testing and the remaining for training. For YouTube dataset, we used the settings given in [5]:

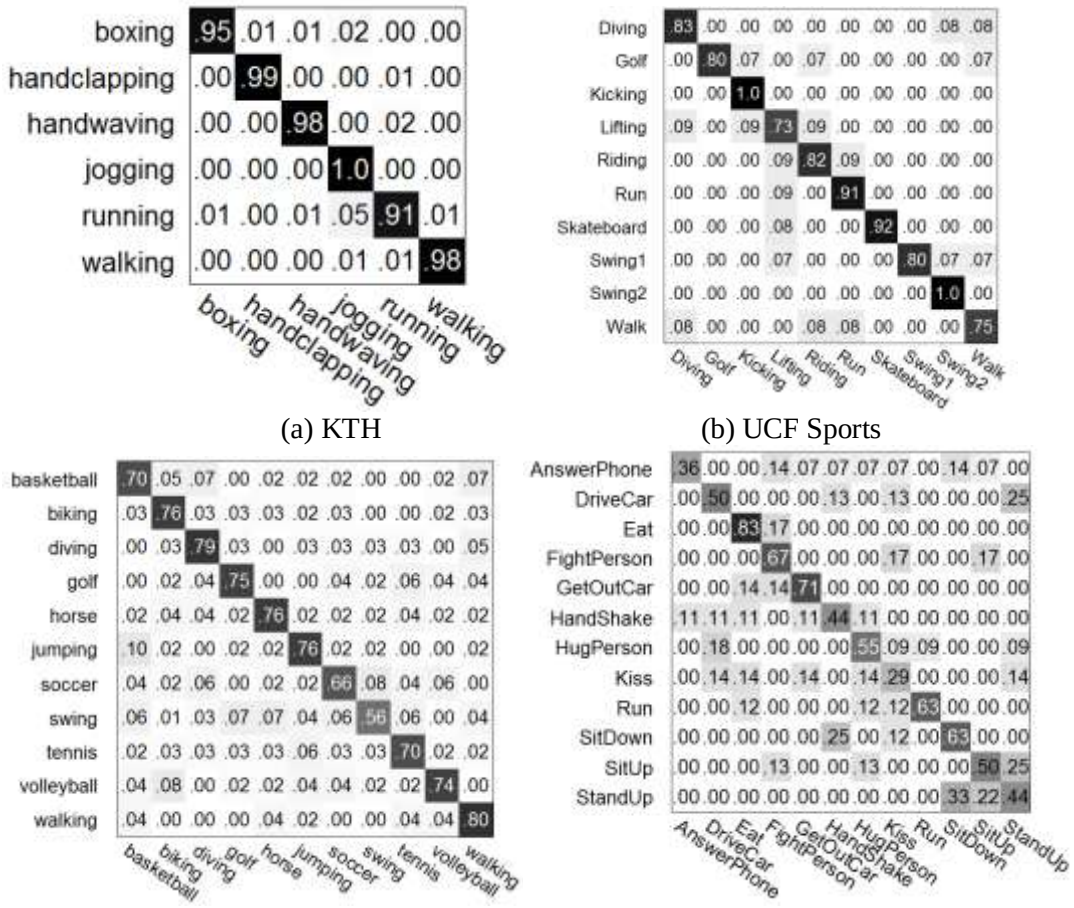
the datasets were divided into 25 subsets, out of which 24 subsets were used for training and the remaining subset was used for testing. For all datasets, we used 48-bin histogram for the motion trajectory descriptor, 32 Fourier coefficients were used in the trajectory shape descriptor, and for the appearance descriptor, we used 128-bin SIFT histogram, and for spatio-temporal interest point, the 640-bin 3D-SIFT descriptors was used for action representation.

Fig 4. From top to bottom : Example of images from KTH, UCF sport ,YouTube and HOHA.

In our method, the feature extraction procedure is computationally expensive, expecially in the training phase. For the KTH, UCF, YouTube and HOHA datasets, the feature extraction takes around 45,26,82,78h respectively.

5.2 Experimental result

In the first experiment, we show the confusion matrices for our method in **Fig. 5.** The major confusion occurs between similar actions such as jogging and running in KTH, and "walk" and "run" in UCF Sports dataset. For HOHA2 dataset, the confusion matrix shows our model is mostly confused by similar action classes, such as "SitUp" with "StandUp" .



(a) KTH (b) UCF Sports

(c) YouTube (d) HOHA

Fig. 5. Confusion matrices of our method on three datasets.

In **Table 1**, we compare the average class accuracy of our method with results reported by other researchers. Compared to the existing approaches, our method shows better performance.

Table 1. Comparison of average class accuracy on action datasets.

	KTH	UCF Sport	YouTube	HOHA2
Wang et al.[25]	91.20%	---	---	---
Rodriguez et al. [4]	88.66%	69.20%	---	---
Liu et al. [15]	94.16%	---	---	---
Lui et al. [35]	97.00%	88.00%	---	---
Sun et al. [7]	---	---	---	47.10%
Wang et al.[33]	92.10%	85.60%	---	47.70%
Yao et al.[34]	92.00%	86.60%	---	---
Bregonzio et al. [1][9]	93.17%	86.90%	64.00%	---
Le et al. [36]	93.80%	86.50%	75.80%	53.30%
Our method	**93.40%**	**85.10%**	**70.30%**	**48.60%**

Excellent result is obtained on the KTH dataset with 93.40% average recognition rate for the six types of actions. This result is better than those obtained by most existing approaches. Note that for KTH and UCF Sport dataset, our results are slightly lower than Lui et al. in [35] which videos are expressed as third order tensors and factorized to a set of tangent spaces. For HOHA2, our approach achieved 48.60% recognition rate. As for the YouTube Dataset, an average recognition rate of 70.30% was obtained. Our results is also slightly lower than that in [36] which used deep learning techniques such as stacking and convolution to learn hierarchical representations and different classifiers.

In **Fig. 6**, we show the effectiveness of affine SIFT for trajectory extraction. It is evident that our trajectory extraction method outperform the methods such as SIFT tracker and L-K tracker used in [7][8][9]. The better result is achieved due to the fully affine-invariance and scale-invariance properties of ASIFT. The improvement is particularly significant for the KTH dataset. **Fig. 6** also shows the pruning process can improve the recognition rate which retains the most relevant trajectories.

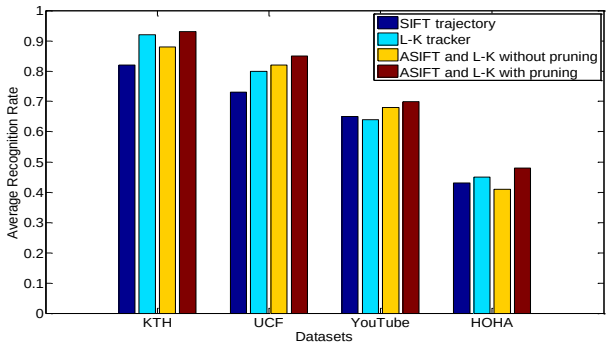


Fig. 6. Comparing effectiveness of different trajectories extraction methods.

We also show the effectiveness of each single descriptor and the fusion of them. It is interesting to note that the trajectory shape descriptor give the better performance among the single descriptor.

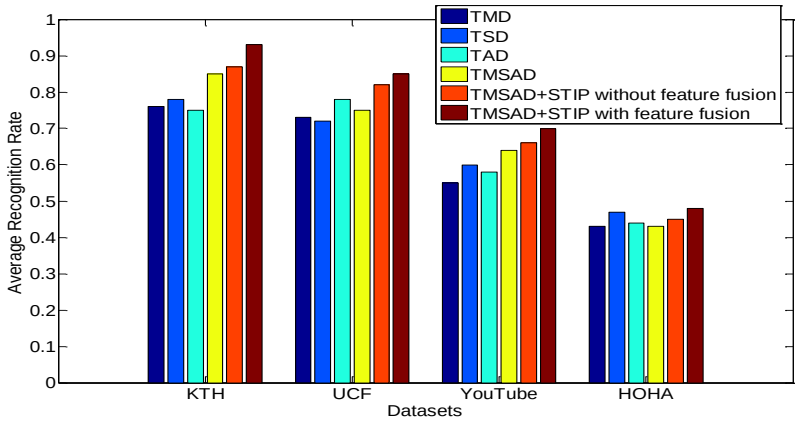
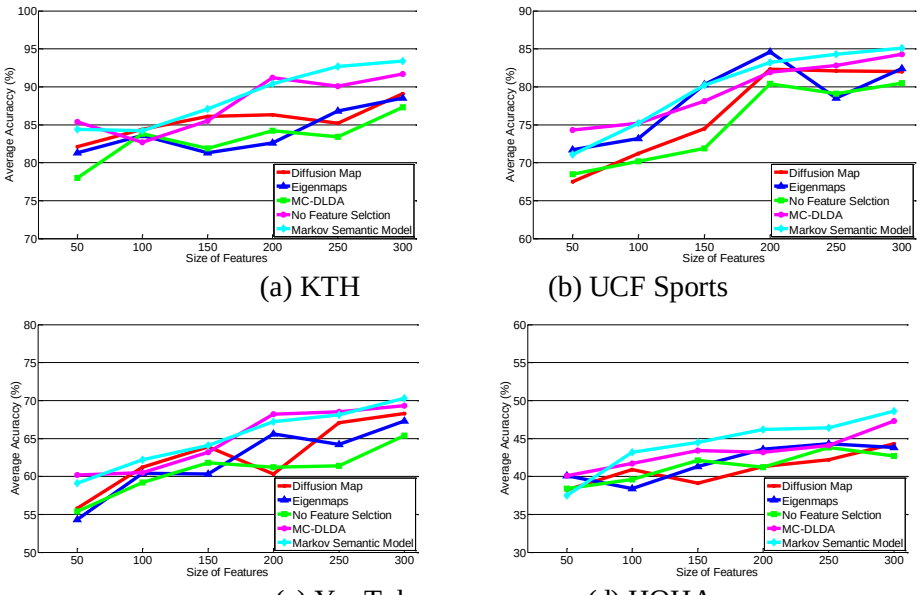


Fig. 7 Evaluation of descriptor performances and effectiveness of feature fusion. TMD: Trajectory Motion Descriptor, TSD: Trajectory Shape Descriptor, TAD: Trajectory Appearance Descriptor, TMSAD: TMD + TSD + TAD, STIP: Spito- Temporal Interest Points

As can be seen in **Fig. 7**, it is evident when the three trajectory descriptors are fused together, action recognition performance is improved. **Fig. 7** also shows that fusion of trajectory based features with local spatiotemporal interest point based features can lead to better performance. The improvement is significant for the YouTube dataset and HOHA2 dataset. This result also demonstrated that different complimentary information can be used simultaneously for action recognition. Following the experiment setting in the [9], we also conducted the experiment that if the trajectories descriptors and spatio-temporal interest features are fused unconditionally without considering the reliability of each type of feature given the camera movements, performance degradation is observed. Mainly because that in the video with camera movements interest points descriptors are less meaningful. If all the interest points features are used for action, they will degrade the recognition performance.

In **Fig. 8** we compare the effectiveness of our feature selection method with a Laplace Eigenmaps proposed in [41], MC-DLDA proposed in [9] and diffusion map proposed in [37]. We firstly embed the midlevel features into a 2000 dimensional space, and then used these selection methods to get the final features for classification. Our feature selection method indeed improves action recognition performance. Note that, the MC-DLDA and our methods got comparative result. This suggests the importance of performing feature selection jointly and collaboratively given highly correlated features. The effect from our MSM feature selection is in particular more significant for YouTube datasets. The clips in this datasets contain more realistic action, thus the original local feature and trajectory feature are less meaningful. Feature selection process based on the semantic words improve the performance.



(a) KTH (b) UCF Sports

(c) YouTube (d) HOHA

Fig. 8. Comparison between different feature selection or reduction methods.

5. Conclusion

We presented a novel framework for recognizing realistic human actions captured from unconstrained environments. We described an action representation scheme by both local interest points and key point trajectories. Different descriptors were extracted for action

representation. A new feature fusion method was utilized for combining trajectory based descriptor with spatio-temporal interest point based descriptors according to the density of the spatiotemporal interest points. Most important, we introduced a novel feature selection approach by formulating a Markov semantic model for feature selection. This model extended the MC-DLDA by different types of dependencies between words produced by short term and long term constraints. The experiment shows that the proposed approach outperforms most existing approaches on the realistic action datasets. The proposed framework achieving comparable results with most existing approaches on action recognition against realistic and unconstrained action recognition datasets.

References

[1] M. Bregonzio, S.Gong and T. Xiang, "Recognising action as clouds of space-time interest points". In *IEEE Conf. Computer Vision and Patt. Recog*, pp.1948-1955, Aug 2009. Article (CrossRef Link).

[2] Fathi and G. Mori, "Action recognition by learning midlevel motion features," In *IEEE Conf. Computer Vision and Patt. Recog*, pp.1-8 , Aug 2008. Article (CrossRef Link).

[3] Z. Zhang, Y. Hu, S. Chan, and L.-T. Chia, "Motion context: A new representation for human action recognition," In *Proc. European Conf. Compute*r Vision, vol.4, pp.817–829, Oct 2008.

[4] M. Rodriguez, J.Ahmed, and M.Shah, "Action MACH: A Spatio-temporal Maximum Average Correlation Height Filter for Action Recognition," In *IEEE Conf. Computer Vision and Patt. Recog*, pp. 1-8, Aug 2008. Article (CrossRef Link).

[5] J. Liu and J. Luo snd M. Shah, "Recognizing realistic actions from videos "in the wild"". In *IEEE Conf. Computer Vision and Patt. Recog*, pp. 1996-2003, Aug 2009. Article (CrossRef Link).

[6] M. Marszałek, I. Laptev, C. Schmid, "Actions in context," *IEEE Conf. Computer Vision and Patt. Recog*, pp. 2929-2936 , Aug 2009. Article (CrossRef Link).

[7] J. Sun, X. Wu, S. Yan et.al., "Hierarchical spatio-temporal context modeling for action recognition," In *IEEE Conf. Computer Vision and Patt. Recog*, pp. 2004-2011, Aug 2009. Article (CrossRef Link).

[8] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," In *IEEE Intl. Conf. Computer Vision*, pp. 104–111, Aug 2009. Article (CrossRef Link).

[9] M. Bregonzio, J. Li, S. Gong and T. Xiang, "Discriminative Topics Modelling for Action Feature Selection and Recognition," In *Proc. Conf. British Machine Vision*, pp. 1-11, Aug 2010.

[10] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing Human Actions: A Local SVM Approach". In *Intl. Conf. Pattern Recognition*, pp. 32-36, Aug2004.

[11] Laptev and T. Lindeberg, "Space-Time Interest Points," In *IEEE Intl. Conf. Computer Vision*, pp. 432-439, Jun 2003. Article (CrossRef Link).

[12] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. "Behavior recognition via sparse spatio-temporal features". In *IEEE Intl. Workshop Visual Surveillance and Performance Evaluation of Tracking and Surveillan*ce,pp. 65-72, Oct 2005. Article (CrossRef Link).

[13] S-F Wong, R. Cipolla, "Extracting spatiotemporal interest points using global information," In *IEEE Intl. Conf. Computer Vision*, pp.1-8, Oct 2007. Article (CrossRef Link).

[14] G. Willems, T. Tuytelaars, L. J. Van Gool, "An efficient dense and scale-invariant spatiotemporal interest point detector," In *Proc. European Conf. Comp*uter Vision, pp. 650-663, Oct 2008.

[15] J. Liu, M. Shah, "Learning human actions via information maximization," In *IEEE Conf. Computer Vision and Patt. Recog*, pp1-8, Aug 2008. Article (CrossRef Link).

[16] J.C. Niebles, F.F. Li, "A hierarchical model of shape and appearance for human action classification," In *IEEE Conf. Computer Vision and Patt. Recog*, pp1-8, Jun 2007. Article (CrossRef Link).

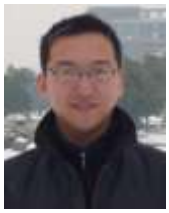
[17] K.G. Derpanis, M. Sizintsev, K.J. Cannons, and R.P. Wildes, "Efficient action spotting based on a spacetime oriented structure representation", In *IEEE Conf. Computer Vision and Patt. Recog*, pp. 1990-1997, Jun 2010. Article (CrossRef Link).

[18] H. Jiang and D. R.Martin, "Finding actions using shape flows," In *Proc. European Conf. Computer Vision*, pp. 278–292, Oct 2008.

[19] P. Matikainen, M. Hebert, and R. Sukthankar, "Trajectons: Action recognition through the motion analysis of tracked features," In *IEEE Conf. Computer Vision and Patt. Recog*, pp.514-521, Jun 2009. Article (CrossRef Link).

[20] J.M. Morel and G.Yu, "ASIFT: A New Framework for Fully Affine Invariant Image Comparison", *SIAM*

Journal on Imaging Sciences, vol.2, no.2, pp. 438-469 , Apr 2009.
[21] M. Steyvers, T. Griffiths, "Probabilistic Topic Models",*Handbook of Latent Semantic A*nalysis, Psychology Press, vol. 427, no.7, pp. 424-440 , Feb 2007.
[22] J.C. Niebles, H-C. Wang, F.F. Li, "Unsupervised learning of human action categories using spatial–temporal words," *International Journal of Computer Vision, vol.*79 no.3, pp. 299–318, Sep 2008.
[23] S-F. Wong, T-K. Kim and R. Cipolla, "Learning motion categories using both semantic and structural information," In *IEEE Conf. Computer Vision and Patt. Recog*, pp.18-23, Jun 2007. Article (CrossRef Link).
[24] J.G. Zhang, S.H. Gong, "Action categorization by structural probabilistic latent semantic analysis," *Computer Vision and Image Understanding*, vol. 114, no.8, pp. 857-864 , May 2010.
[25] Y. Wang, G. Mori, "Human Action Recognition by Semi-latent Topic," Models. *IEEE Trans. Pattern Anal. Mach. Intel*, vol. 31 no.10, pp.1762-1774 , Oct 2009. Article (CrossRef Link).
[26] [26] T. Hospedales, S. Gong and T. Xiang, "A Markov Clustering Topic Model for Mining Behaviour in Video," In *IEEE Intl. Conf. Computer Vision*, pp.1165-1172 , Oct 2009. Article (CrossRef Link).
[27] D. M. Blei, M. I. Jordan and A. Y. Ng, and Jafferty, "Latent Dirichlet allocation." *Journal of Machine Learning Research*, vol.3, no.4, pp. 993–1022, Jan 2003.
[28] P. Scovanner, S. Ali, M. Shah, "A 3-dimensional SIFT descriptor and its application to action recognition", In *Intl Conf. Multimedia*, pp.357–360, Sep 2007.
[29] T. Griffiths, M. Steyvers, D. M. Blei, J. B. Tenenbaum. "Integrating topics and syntax". *Advances in Neural Information Processing Systems* vol.17 no.17, pp. 537-544 , Dec 2005.
[30] T. Griffiths, M. Steyvers, "Finding scientific topics". *Proceedings of the National Academy of Sciences*, vol. 101, no.1, pp.5228–5235 , Apr 2004.
[31] T. Minka, J. Lafferty, "Expectation-propagation for the generative aspect model," In *Proc. Conf. Uncertainty in Artificial Inte*lligence, pp.352-359 , Aug 2002.
[32] S. Maji, A.C. Berg, J. Malik, "Classification using intersection kernel support vector machines is efficient," In *IEEE Conf. Computer Vision and Patt. Recog*, pp.1-8, Jun 2008. Article (CrossRef Link).
[33] H. Wang, M. Muneeb Ullah, A. Kläser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," In *Proc. Conf. British Machine Vision*, pp.1-10 , Sep 2009.
[34] Yao, J. Gall, and L. V. Gool, "A hough transform-based voting framework for action recognition". In *IEEE Conf. Computer Vision and Pat*t. Recog, pp. 2061 - 2068, Jun 2010. Article (CrossRef Link).
[35] Y. M. Lui and R. Beveridge, "Tangent bundle for human action recognition," In *Proc. Conf. Automatic Face and Gesture Recognition*, pp.97-102 , Mar 2011.
[36] Q.V. Le, W.Y. Zou, S.Y. Yeung, A.Y. Ng. "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," In *IEEE Conf. Computer Vision and Patt. Rec*og, pp.3361 - 3368, Jun 2011. Article (CrossRef Link).
[37] J. Liu, Y. Yang and M. Shah, "Learning semantic visual vocabularies using diffusion distance," *IEEE Conf.* Computer Vision and Patt. Recog, pp.461 - 468, Jun 2009. Article (CrossRef Link).
[38] D. Zhang and G. Lu, "Study and evaluation of different fourier methods for Image Retrieval", *Image and Vision Computing*, vol.23, no.1 pp. 33-49, Ja 2005.
[39] H. Wang and Alexander et al. Action Recognition by Dense Trajectories,In *IEEE Conf. Computer Vision and Patt. Recog*, pp. 3169-3176, Aug 2011. Article (CrossRef Link).
[40] J. Liu, J. Yang. "Action recognition using spatiotemporal features and hybrid generative/discriminative models". *Journal of Electronic Imaging*. vol.21 no.2, pp.1-11 Apr 2012.
[41] M. Belkin and P. Niyogi, "Laplacian Eigenmaps for dimensionality reduction and data representation". *Neural Computation*, vol.15, no. 6, pp.1373-1396, Jun 2003.



Jia Liu is currently a PhD candidate at the Institute of Image Processing and Pattern Recognition at Shanghai Jiao Tong University, Shanghai, China. He received her BE degree and MS degree from the Engineering College of Armed Police Force, Xi'an, China. His research interests are computer vision, machine learning, action recognition..

Xiaonian Wang is currently an MS candidate at the Institute of Image Processing andPattern Recognition at Shanghai Jiao Tong University, Shanghai, China. His research interests are computer vision and visual surveillance.

Jie Yang received his PhD in computer science from Hamburg University, Germany. He is now a professor at Shanghai Jiao Tong University (SJTU) and dean of the Institute of Image Processing and Pattern Recognition at SJTU. He has performed more than 20 national scientific research projects in image processing, pattern recognition, data mining, and artificial intelligence.

Tianyu Li is currently an MS candidate at the Institute of Image Processing and Pattern Recognition at Shanghai Jiao Tong University, Shanghai, China. His research interests are object tracking and computer vision.