Hierarchical Filtered Motion for Action Recognition in Crowded Videos

YingLi Tian, *Senior Member*, *IEEE,* Liangliang Cao, *Student Member, IEEE,* Zicheng Liu, *Senior Member, IEEE,* and Zhengyou Zhang, *Fellow, IEEE*

Abstract—**Action recognition with cluttered and moving background is a challenging problem. One main difficulty lies in the fact that the motion field in an action region is contaminated by the background motions. We propose a Hierarchical Filtered Motion (HFM) method to recognize actions in crowded videos by using Motion History Image (MHI) as basic representations of motion due to its robustness and efficiency. First, we detect interest points as the 2D Harris corners with recent motion, e.g. locations with high intensities in MHI. Then a global spatial motion smoothing filter is applied to the gradients of MHI to eliminate isolated unreliable or noisy motions. At each interest point, a local motion field filter is applied to the smoothed gradients of MHI by computing a structure proximity between any pixel in the local region and the interest point. Thus the motion at a pixel is enhanced or weakened based on its structure proximity with the interest point. To validate its effectiveness, we characterize the spatial and temporal features by Histograms of Oriented Gradient (HOG) in the intensity image and MHI respectively and use a Gaussian Mixture Model (GMM) based classifier for action recognition. The performance of the proposed approach achieves the state-of-the-art results on KTH dataset which has clean background. More importantly, we perform cross dataset action classification and detection experiments where KTH dataset is used for training while MSR Action Dataset II, which consists of crowded videos with people moving in the background, is used for testing. Our experiments show that the proposed hierarchical filtered motion method significantly outperforms existing techniques.**

Index Terms—**Action classification, action detection, crowded videos, hierarchical filtered motion, Motion History Image.**

I. INTRODUCTION

ACTION recognition with cluttered and moving background is a challenging problem which captures increasing interests [1-2, 4-6, 8-12, 14-28, 30-36, 42-49, 51, 53-55, 59]. One main difficulty lies in the fact that the motion field in an action region is contaminated by the background motions. The goal of our work is to recognize human actions in dynamic and crowded environments by using the action models which are trained on data with clean background. To achieve the goal, we need to develop techniques to obtain stable and generalized features and feature descriptors that are able to characterize actions but insensitive to cluttered background motions.

Recent work on action recognition in realistic videos has focused on approaches based on bag-of-spatio-temporal-features [22, 32, 55], spatio-temporal shapes [14], and histories of tracked keypoints [32, 41]. Compared with classification task, action detection is more challenging. There are only a few works devoted to action detection task [5, 6, 11, 18, 57, 53]. Laptev *et al.* [22, 28] used local spatio-temporal invariant points (STIPs) [21], space-time pyramids, local spatial-temporal descriptors (HOG/HOF) [7, 40], and multichannel non-linear SVMs for realistic actions in movies. Yuan *et al.* [55] employed the same features (STIPs) and descriptors (HOG/HOF) and proposed a discriminative subvolume search for efficient action detection by using a Nearest Neighbor based classifier. Ke *et al.* [18] proposed a method to detect event in crowded videos by combining spatio-temporal shapes with a flow descriptor. Sun *et al.* [44] modeled the spatio-temporal information for action recognition in realistic datasets at 3 levels: point-level, intra-trajectory level, and inter-trajectory level. The trajectories are extracted based on matching the SIFT salient points over consecutive frames. Similarly, Messing *et al.* [32] proposed a system for action recognition by using the velocity histories of tracked keypoints which are extracted by Kanade-Lucas-Tomasi (KLT) feature trackers, and used a generative mixture model to learn and classify actions. They also augmented other features such as position, appearance, color, etc. to improve the recognition accuracy. Junejo *et al.* [16] attempted to recognize human actions under different views using temporal self-similarities. Yin and Meng [54] proposed a method to learn the shapes of space-time feature neighborhoods for each action category. Surveys of video event understanding and human motion analysis can be found in papers [15, 23]. While paper [15] reviews the recent advances in view-invariant human motion analysis, paper [23] summarizes the most recent methods for automatic interpretation of semantic occurrences in video. Despite promising results are achieved by the state-of-the-art work,

Manuscript received August 6, 2010, revised December 20, 2010, accepted April 24, 2011. This work was supported in part by NSF grant IIS-0957016 and Microsoft Research.

YingLi Tian is with the City College, City University of New York, New York, NY 10031 USA (phone: 212-650-7046; fax: 212-650-8249; e-mail: ytian@ccny.cuny.edu). Prior to joining the City College in September 2008, she was with IBM T.J. Watson Research Center, Yorktown Heights, NY 10598 USA

Liangliang Cao is with Beckman Institute and Coordinate Science Lab, Dept. ECE, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: cao4@illinois.edu).

Zicheng Liu is with Microsoft Research, Redmond WA 98052 USA (e-mail: zliu@microsoft.com).

Zhengyou Zhang is with Microsoft Research, Redmond WA 98052 USA (e-mail: zhang@microsoft.com).

more robust methods are needed to handle cluttered background motions due to the following difficulties: 1) there is no mechanism to distinguish action motions and background motions in existing local spatio-temporal interest point detectors and descriptors, and 2) the trajectories of keypoints cannot be reliably tracked in crowded videos.

In this paper, we propose an efficient hierarchical filtering technique to extract motion information and reduce the distracting motions caused by the background moving objects. Figure 1 shows the framework of our method for action recognition in crowded videos. Instead of using spatio-temporal invariant points, we extract spatial and temporal information separately. The spatial information is extracted as 2D Harris corners in original image, and the temporal information is obtained from Motion History Image (MHI) [4, 8] which is based on frame differencing. Using MHI allows us to avoid unreliable keypoint tracking in crowded videos. The pixels in MHI with brighter intensities which represent the moving objects with more recent motion are formed as a template. We combine this motion template and the extracted 2D Harris corners for interest point detection. Only those corners with the most recent motion are selected as interest points. We observe that an isolated motion direction of a pixel compared to its neighbor pixels is often a distracting motion or a noise. To remove the isolated distracting motions, we first apply a global spatial motion smoothing filter to the gradients of MHI. At each interest point, a local motion field filter is applied by computing a structure proximity between any pixel in the local region and the interest point. Thus the motion at a pixel is enhanced or weakened based on its structure proximity with the interest point. To characterize the temporal features, we present a new temporal feature descriptor – Histograms of Oriented Gradient in Motion History Image (HOG-MHI). The spatial features are modeled by HOG in the intensity image as the existing work. The feature vectors which contain both HOG (spatial features) and HOG-MHI (temporal features) are modeled by a Gaussian Mixture Model (GMM) based classifier for action recognition.

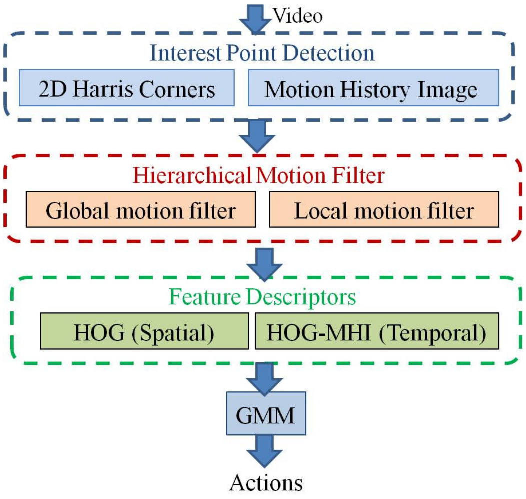


Fig. 1. The framework of proposed method for action recognition in crowded videos. The components with shadow indicate our main contributions.

The performance of the proposed approach achieves the state-of-the-art results on standard KTH dataset which has clean background at 93.9% accuracy. In order to validate the efficiency, effectiveness, and generalizability of the proposed method to handle cluttered background, we perform action recognition and cross-dataset test on MSR action dataset II which consists of three classes of actions (handclapping, handwaving, and boxing as in KTH dataset) with people moving around or vehicles driving by in the background. We demonstrate that our method is of high computational efficiency for real-time action recognition and significantly outperforms existing techniques in crowded videos.

II. Hierarchical Motion Field Filters for Action Detection

A. Interest Point Detection

Motion History Image (MHI): MHI is a real-time motion template that temporally layers consecutive image differences into a static image template [4, 8]. Pixel intensity is a function of the motion history at that location, where brighter values correspond to more recent motion. The directional motion information can be measured directly from the intensity gradients in the MHI. Compare to optical flow, gradients in the MHI are more efficient to compute. It is also more robust due to the fact that the motion information in MHI is mainly along the contours of the moving objects. Thus, unwanted motion in the interior regions of object contours is ignored.

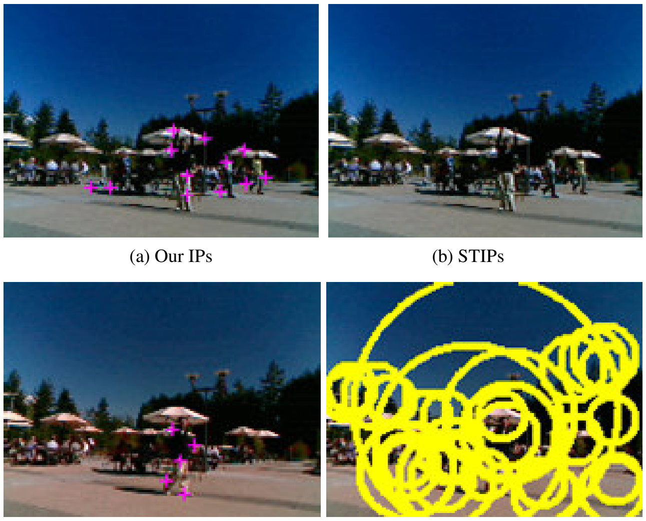
To generate a MHI, we use a simple replacement and decay operator as in paper [4]. At location *(x, y)* and time *t*, the intensity of $MHI_\tau(x,y,t)$ is calculated:

$$MHI_\tau(x,y,t) = \begin{cases} \tau, & if\ D(x,y,t)=1 \\ max(0, MHI_\tau(x,y,t-1)-1), & otherwise \end{cases} \quad (1)$$

where $D(x,y,t)$ is a binary image of differences between frames and τ is the maximum duration of motion. We set τ as 20 in our system based on experiments. The MHI image is then scaled to a grayscale image with maximum intensity 255 for pixels with the most recent motion.

Interest Point Detection: Sparse selection of spatio-temporal interest points has been successfully used for action recognition [6, 9, 21, 22, 30, 35, 42, 55]. Laptev *et al.* developed a nice mathematic framework to find pixels with significant variations in both spatial and temporal directions [21]. However, the interest points detected by their approach are in practice too sparse to characterize well the motion features. Dollar *et al.* proposed to detect the interest points by extracting the maximum response of Gabor filter [9]. The limitation of the approach in [9] is that the filtering parameters are sensitive in complex scenes and the detected interest points are heavily affected by the cluttered background and foreground occlusions.

We have tested STIPs [22] for the videos with cluttered background by using different parameter settings of scales and observed that there are not enough interest points in action regions (Figure 2(b)). In some sequences with large lighting changes, many STIPs are extracted on the background as shown in Figure 2(d). To overcome the above limitation, our interest point detection is based on detecting corners in images (2D Harris Corner Detection [13]) and combining the temporal information which are obtained from MHI. Harris Corner detection is stable to different scales and insensitive to lighting changes. Here, we use MHI as motion mask to remove the corners in the static background. Only the corners with more recent motion (intensity in MHI > threshold) are selected as interest points. Figure 3 shows several examples of interest point detection from KTH dataset. The top rows show the detected 2D Harris corners in the images. Bottom rows show the MHIs and the interest points. Examples on MSR Action Dataset II with cluttered background can be found in Figure 4. In Figure 2, we demonstrate the effectiveness and robustness of our interest point detection method (Figure 2(a) and 2(c)) by comparing with the STIP detection method (Figure 2(b) and 2(d)) which was developed by Laptev *et al.* [11]. Note that essential STIPs are missed on action regions (Figure 2(b)) and many false STIPs are detected on background regions due to lighting changes (Figure 2(d)). Our method is more robust to lighting changes.


(a) Our IPs (b) STIPs

(c) Our IPs (d) STIPs

Fig. 2: Examples of interest point detection by our method and STIP detection of Laptev *et al.* [22] in a video with cluttered background and lighting changes. (a) Interest points are detected on moving people by our method; (b) no STIPs are detected by [22]; (c) our interest point detection is insensitive to lighting changes; and (d) false STIPs are detected on background regions [22].

B. Hierarchical Filtered Motion Field for Action Motion Enhancement

The motion template based on MHI has been used for action recognition by assuming the action of interest is well segmented from the background. Bobick and Davis [4] used the motion template to recognize many types of aerobics exercises. Weinland *et al.* [49] extended MHI to Motion-History Volumes for free-viewpoint action recognition in the setting of multiple calibrated cameras and background subtracted. Meng *et al.* [31] proposed a method for action recognition by using histogram of MHI and the Haar wavelet transform [37, 38] of MHI and achieved 71% classification accuracy on the KTH dataset.

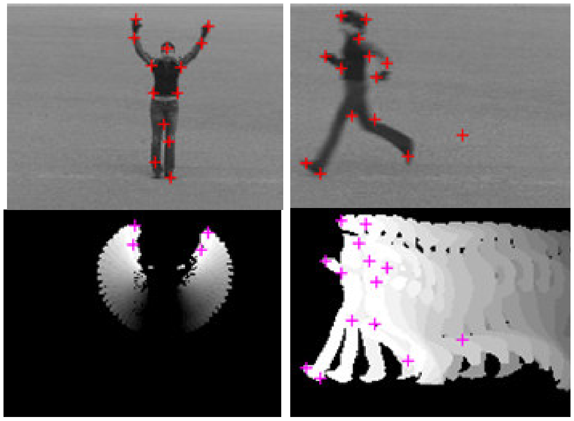
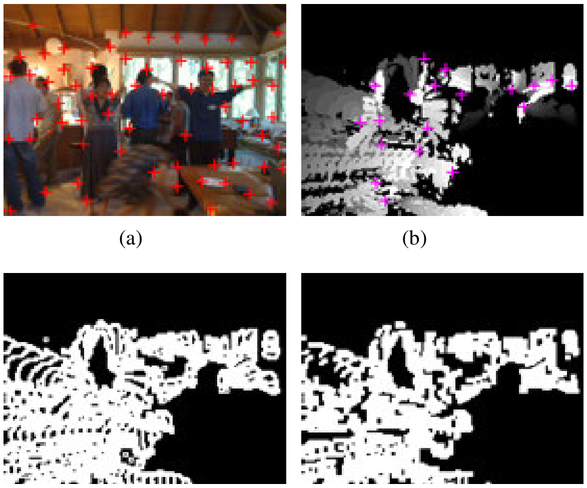

(a) hand waving (b) running

Fig. 3: Examples of interest point detection by applying the MHI as a motion mask on 2D Harris corners. The detected 2D Harris corners are displayed in the original images (red "+"). The detected interest points are displayed in the MHI (pink "+".)


(a) (b)

(c) (d)

Fig. 4: Global filtered motion filed to eliminate isolated distracting motions. (a) Original image with 2D Harris corners (red "+"); (b) MHI with detected interest points (pink "+"); (c) the binary image of the intensity gradients of MHI (MGI); (d) smoothed gradients of MHI.

In order to handle cluttered background, we propose a hierarchical filtered motion field technique based on Motion Gradient Image (MGI). The MGI is the intensity gradients of MHI which directly yield the motion orientation. Note that the magnitudes of the MHI gradients are not meaningful. Although it is impossible to distinguish the action motions from the background motions without using high-level information, we still can reduce noisy motions and enhance the action motions based on the following observations: 1) an isolated motion direction of a pixel compared to its neighbor pixels is often a distracting motion or a noisy motion, and 2) at each interest point, the motion regions which are closer to the interest point contribute more to the object which the interest point belongs to.

Global Filtered Motion Field: In our approach, we first apply a motion smoothing step at the MGI to remove the isolated motion directions by morphological operations to obtain a global filtered motion field – smoothed gradients of MHI. We show one example in Figure 4. Figure 4(a) shows one frame of the original image with 2D Harris corners (red "+"). Figure 4(b) shows the MHI of the same frame with detected interest points (pink "+"). Figure 4(c) displays the binary image of the intensity gradients of MHI (MGI), and Figure 4(d) is the smoothed gradients of MHI which removed the isolated distracting motions. To be prepared for local filtered motion field processing, we decompose the smoothed gradients of MHI as a number of layers with different motion directions. Figure 5 illustrates an 8-bin-layer representation of a binary image of the smoothed gradients of MHI. As shown in Figure 7, for a total 8 bin HOG-MHI, the motion directions for each bin fall in the range of $n \pm 22.5°$ ($n = 1, 2, \ldots 8$).

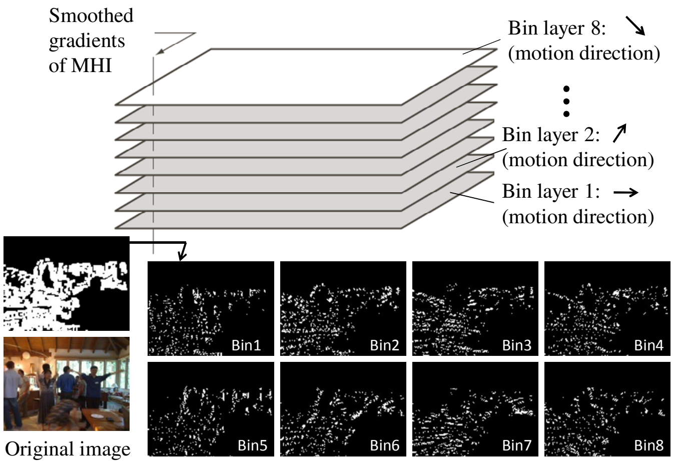

Smoothed gradients of MHI
Bin layer 8: ↘ (motion direction)
Bin layer 2: ↗ (motion direction)
Bin layer 1: → (motion direction)
Bin1
Bin2
Bin3
Bin4
Bin5
Bin6
Bin7
Bin8
Original image

Fig. 5: Bin-layer representation of a binary image of the smoothed gradients of MHI for 8 bins HOG-MHI.

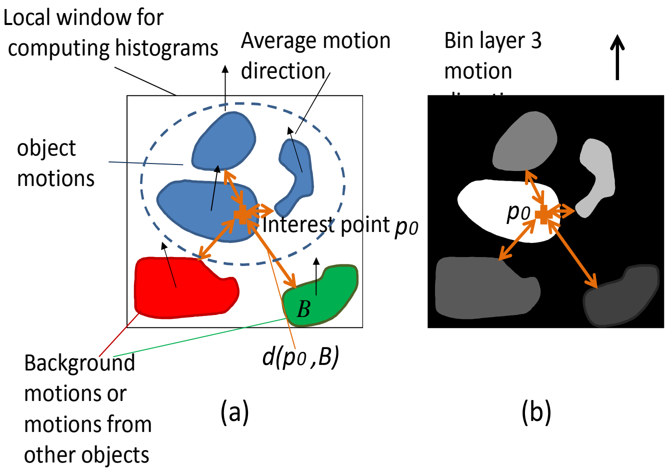


Fig. 6. Local motion field filtering. (a) Motion blobs in a local window of an interest point; (b) Plot of the structure proximity map.

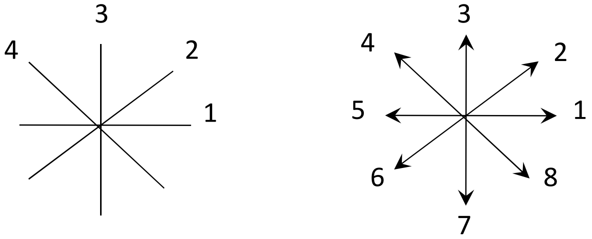
Local Filtered Motion Field: At each interest point, we apply a local filtered motion field by computing a structure proximity between the pixels in the local region and the interest point on each bin-layer of the smoothed gradients of MHI. Here the local region is the window for calculating HOG-MHI. A connect component operation is performed to obtain motion blobs. Figure 6(a) illustrates the blobs of bin-layer 3, the motion blobs with shorter distances to the interest point in the local region are more likely to represent the motion of the object which the interest point belongs to. Thus the motions at these blobs (blobs in blue color) should be enhanced. On the other hand, the blobs with longer distances to the interest point most likely belong to other objects (blobs in red and green colors). Thus the motions at those blobs should be weakened. Let p_0 denote the interest point. Let B denote a blob. Denote $d(p_0, B)$ to be the minimum distance between p_0 and all the points in B, that is,

$$d(p_0, B) = \min_{p \in B} d(p_0, p)$$

Denote W_x, W_y to be the size of the window. Then the maximum distance between p_0 and any points in the window is $\sqrt{W_x^2 + W_y^2}/2$. For any pixel $p \in B$, we define its structure proximity to interest point p_0 as

$$s(p) = 1 - \frac{2d(p_0, B)}{\sqrt{W_x^2 + W_y^2}} \qquad (2)$$

Note that $s(p)$ is a value between 0 and 1. If a pixel does not belong to any blobs, we define its structure proximity to be 0. Figure 6(b) shows a plot of the structure proximity map where brighter intensity values indicate larger structure proximity values. The structure proximity values are used to normalize motion histograms in HOG-MHI calculation.



(a) HOG (b) HOG-MHI

Fig. 7. HOG/HOG-MHI descriptors. (a) No directions for appearance features, and (b) directions are considered for motion features.

C. HOG and HOG-MHI Feature Descriptor

HOG feature descriptors have been widely used in human detection and action recognition [7, 22, 28, 40, 55]. In our system, the local appearance and motion features are characterized by grids of Histograms of Oriented Gradient (HOG) in the neighborhood with a window size (W_x, W_y) at each interest point in the intensity image and MHI respectively. The window is further subdivided into a (n_x, n_y) grid of patches. Normalized histograms of all the patches are concatenated into HOG (for appearance features in the intensity image) and HOG-MHI (for motion features in the MHI) descriptor vectors as the input of the classifier for action recognition. As shown in Figure 7, the calculations of HOG and HOG-MHI are different. We compute HOG without considering the directions to make it more robust to appearance changes. However, for HOG-MHI computation, the performance of action recognition decreases without considering directions since directions are important to describe motion features. In our experiments, we set $n_x, n_y = 3$ and use 6 bins for HOG in the intensity image and 8 bins for HOG-MHI in the MHI image). For each interest point, the HOG (with dimension of 54) and HOG-MHI features (with dimension of 72) are concatenated into one feature vector for action classification.

To handle scale variations, a multi-scale process at each interest point can be applied by using different patch sizes or by using same patch size on different scale images. However, the multi-scale process will heavily increase the size of the feature vector for training and testing. For example, the size of the feature vector will be tripled for three scales. Thus, instead of performing a multi-scale process at each interest point, we use randomly selected window sizes between W_{min} and W_{max}. The size of each window is calculated by $W_x = kn_x$ and $W_y = kn_y$ where k *is* randomly chosen to make sure the values of W_x, W_y are in between W_{min} (minimum window size) and W_{max} (maximum window size). In our experiments, we set $W_{min} = 24$, $W_{max} = 48$, and $n_x, n_y = 3$. Our experiments demonstrate that using randomly selected window sizes handles scale variations very well and achieves better results than using fixed set of scales.

As we mentioned in Section *II-B*, the magnitude of the intensity gradients of the MHI is not meaningful. To normalize the histograms of MGI, we use the structure proximity values instead of the magnitudes at each patch.

III. Gaussian Mixture Model (GMM) for Action Classification

The simplest model to model the feature descriptor is normal distribution. However, a single normal distribution is not enough to characterize the complex nature of rich descriptors. We employ a Gaussian Mixture Model (GMM), which is known to have the ability to model any given probability distribution function when the number of mixture component is large. Given a K component GMM, the probability of a patch x is

$$\Pr(x \mid \Omega) = \sum_{k=1}^{K} w_k \mathrm{N}\ (x; \mu_k, \Sigma_k) \qquad (3)$$

where $\mathrm{N}\ (x; \mu_k, \Sigma_k)$ denotes the normal distribution with mean μ_k and variance Σ_k :

$$\mathrm{N}\ (x; \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{d/2} \mid \Sigma_k \mid^{1/2}} \mathrm{e}^{-\frac{1}{2}(x-\mu_k)^T (\Sigma_k)^{-1} (x-\mu_k)}$$

where $x \in R^d$, $\mu_k \in R^d$, and $\sum_k \in R^d \times R^d$. The mixture weight w_k satisfies the constraint

$$\sum_{k=1}^{K} w_k = 1$$

The set of all the parameters of GMM model is denoted as $\Omega = \{w_k, \mu_k, \Sigma_k\}$, $1 \le k \le K$.

Although the general mode in equation (3) supports full covariance matrices, in practice a diagonal covariance matrix is enough for most of the tasks. Furthermore, diagonal matrix GMMs are more computational efficient and robust compared with full matrix GMM.

The advantages of using a GMM are that it is computationally inexpensive, and it is based on a well-understood statistical model. With GMM, we can clearly estimate the probability that each patch belongs to the background or the action of interests, which can be used to distinguish patches of actions of different categories and the background. Suppose there are C categories of actions with the parameter of Ω^1, Ω^2, …, Ω^C. Each category corresponds to a GMM with K components $\Omega^c = \{w_k^c, \mu_k^c, \Sigma_k^c\}$.

The parameters can be estimated using maximum likelihood estimation. For example, for the c *th* category, we first collect all the patches X^c associated with action c , and then estimate Ω^c via the maximum estimation of

$$\max_{\Omega^C} L^c = \max_{\Omega^C} \sum_{x_i \in X^C} \log \Pr(x_i \mid \Omega^c)$$

This can be solved by EM algorithm, which is an iterative method alternating between performing an expectation step (E-step) and a maximization step (M-step). In the E-step, we estimate the posterior probability for each sample. In the M-step, we update $\{w_k, \mu_k, \Sigma_k\}$ based on the posterior probability. The E-step and M-step are repeated until converge.

A straightforward way to train these models is to train Ω^1, $\Omega^2, \ldots, \Omega^C$ separately. However, Reynolds *et al.* [41] showed it is more effective to obtain Ω^1, Ω^2, …, Ω^C coherently by the use of a universal background model. This observation has also been validated by Yan *et al.* [52]. Following this approach, we first train an action-independent background model Ω^0 based on all the patch features $x \in X^{all}$. Then we adapt Ω^1, Ω^2, …, Ω^C from Ω^0 by changing the $\{\mu_k^c\}$ in the following way

$$p_{ik}^c = \frac{w_k \mathrm{N}\ (x_i; \mu_k^0, \Sigma_k^0)}{\sum_{k'=1}^{K} w_{k'} \mathrm{N}\ (x_i; \mu_{k'}^0, \Sigma_{k'}^0)} \text{ for } x_i \in X^c \qquad (4)$$

$$\mu_k^c = \frac{1}{n^c} \sum_{x_i \in X^c} p_{ik}^c x_i$$

Theoretically we could update w_k^c and Σ_k^c as well as μ_k^c. However, in practice we force them to be the same as the background model. Compared with the approach which trains action model separately, the use of background model is more computational efficient and leads to a good alignment of different action models over different components, which makes the recognition more accurate.

After obtaining the GMM parameters Ω^1, Ω^2, …, Ω^C, we can easily classify a new video clip according to the action category. Let V denote the collection of patch descriptors in a video clip, we can estimate the action category by

$$c^{*} = \arg\max_{c} \sum_{x \in V} \log \Pr(x \mid \Omega^{c}) \qquad (5)$$

IV. Experiments and Discussion

A. Databases

The KTH dataset [41] was used as a standard benchmark for action recognition. It was recorded in four controlled environments with clean background (indoors, outdoors, outdoors with scale variation, outdoors with different clothes.) The dataset contains about 600 video sequences of 25 subjects performing six categories of actions: boxing, hand clapping, hand waving, jogging, walking, and running. The video resolution is 160x120.

To validate the efficiency and robustness of our method in crowded videos, we perform cross dataset testing on MSR Action Dataset II [60] which contains three types of actions selected from KTH: boxing, hand clapping, and hand waving. There are in total 54 video sequences with cluttered background in different environments such as cafeterias, home, and street. The dataset includes in total 81 boxing, 51 hand clapping, and 71 hand waving actions. Some actions are simultaneously performed by multiple people. The video resolution is 320x240.

B. Action Classification Results on KTH dataset

To make the results comparable, we apply exactly the same experimental setting of KTH dataset as in [9, 22, 33, 55]. Among the 25 persons, we use16 persons (1528 sequences) for training and the other 9 persons (863 sequences) for testing.

In the experiments, we investigated the effects of 1) the motion duration for MHI calculation, 2) use only HOG, HOG-MHI, and both of them, 3) the number of bins for appearance (HOG) and motion (HOG-MHI). Our method is insensitive to the parameter of motion duration. For the 6 categories of actions, the results are relatively stable when the motion duration changes from 10 – 25 frames. As shown in Table III, the recognition accuracy changes from 89.2% to 93.9%. The best results are achieved when we use both HOG (with 6 bins without orientation) and HOG-MHI (with 8 bins with orientation) with the motion duration at 20 frames. The confusion matrix of the action recognition for the KTH dataset is presented in Table I. The average accuracy is 93.9%. Most of the mis-recognitions are from the confusion between jogging and running.

Table II shows that our method achieves the state-of-the-art results by comparing with previous work on the KTH dataset [9, 17, 19, 22, 33, 54, 55]. To understand the effects of the features and the classifier, we perform the experiment by using the same features (STIPs) and descriptors (HOG/HOF) as in paper [22, 55] but using GMM as the classifier, and a slightly better accuracy is achieved than using the proposed Hierarchical Filtered Motion features and descriptors on the KTH dataset. This shows that Hierarchical Filtered Motion features and feature descriptors are comparable with the state of the art methods. Further, to understand the effects of different classification methods, we also applied STIP based features with GMM-based classifier. We observe that the GMM based classier is slightly better than SVM [22] and NBMIM [55] classifiers. We also test the proposed hierarchical filtered motion field on KTH dataset, and there is no obvious difference between with (93.6%) and without (93.9%) using the hierarchical filtered motion. This is not surprising since the background of KTH dataset is clean.

TABLE I
Confusion matrix of action recognition on the KTH action dataset (6 actions with clean background)

	boxing	andclapping	handwaving	jogging	running	walking
boxing	**134.0**	4.0	2.0			4.0
handclapping	3.0	**138.0**	3.0			
handwaving		1.0	**143.0**			
jogging				**134.0**	10.0	
running				20.0	**124.0**	
walking				6.0		**138.0**

TABLE II
Comparison with the state-of-the art results on the KTH action dataset (6 actions with clean background)

Method	Accuracy
Dollar *et al.* [9]	80.7%
Yin *et al.* [54]	82%
Kaaniche *et al.* [17]	90.57%
Laptev *et al.* [22]	91.8%
Mikolajczyk e*t al.* [33]	93.2%
Yuan *et al.* [55]	93.3%
Kovashka *et al.*[19]	94.53%
STIP + HOG/HOF + GMM	**94.5%**
2D corners + IP Detection + HOG/HOG-MHI + GMM	**93.9%**
2D corners + IP Detection + Hierarchical motion filter + HOG/HOG-MHI + GMM	**93.6%**

TABLE III
Results of using different MHI durations for our method (2D corners + IP Detection + HOG/HOG-MHI + GMM)

MHI duration (frames)	Accuracy
10	89.2%
15	93.1%
20	93.9%
25	92.5%

C. Cross-dataset Action Classification Results

Cross-dataset classification and detection for actions is important for real surveillance applications. Conventional classifier and detector usually are trained from labeled examples and assume the testing samples are generated from the same distribution. For a new dataset with a different distribution from the training dataset, a new training process is needed, which requires large amount of training labels from the new dataset. Here, we perform cross-dataset for action classification to demonstrate the robustness and generalizability of the proposed Hierarchical Filtered Motion features.

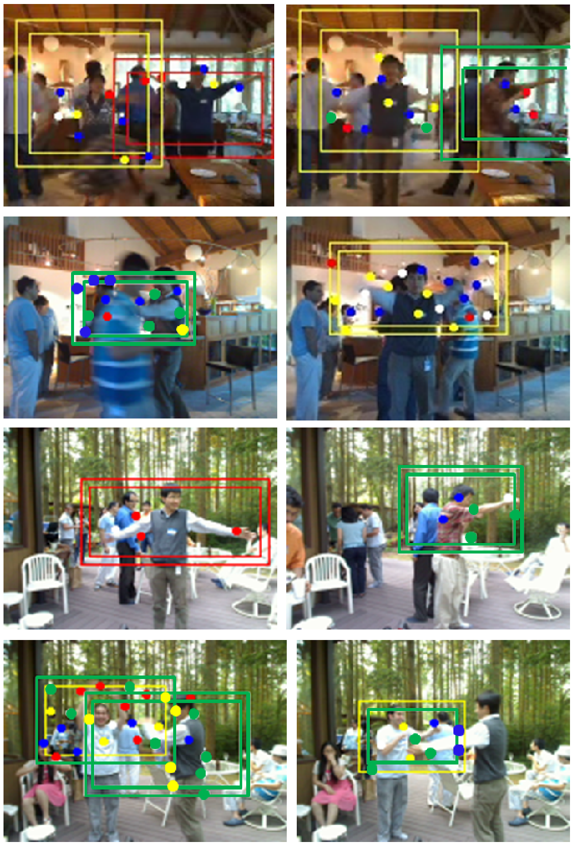


Fig. 8. Examples of action recognition results on MSR Action Dataset II. The type of actions is coded by different colors: yellow – hand waving, red – hand clapping, and green – boxing. For each action, the inner box shows the extracted action subvolume obtained from ground truth and the outer box shows the action recognition results based on the classification scores in the action subvolume. The action is correctly recognized if the colors of the inner and outer boxes are identical.

MSR Action Dataset II contains 3 action categories: boxing, hand clapping, and hand waving, which are the same types as the KTH dataset. We test Hierarchical Filtered Motion features for action classification in the cross-dataset scenarios. In order to directly use the GMM model trained from the KTH dataset for 6 actions to recognize the actions on MSR Action Dataset II, we first downsample the video sequences to the same resolution as the KTH dataset (160x120). Then the exact same set parameters are used to extract interest points and calculate HOG and HOG-MHI. The action recognition is performed on the extracted action subvolumes based on the ground truth labels. Some examples of action recognition are shown in Figure 8. Note that the extracted action subvolumes contain background motions. As shown in the left image of the second row, the action is correctly recognized even with short time full occlusion. The last row displays two examples of the actions which are not correctly classified due to different actions are performed in the same region for the whole action period.

The detailed results are presented in Table IV. Among the 203 actions, we achieve 78.8% recognition rate. Although we directly use the classifier trained on the KTH dataset with 6 types of actions, for the testing on MSR Action Dataset II, all the errors are from the mis-classification between the 3 actions of upper body. There is no confusion between the upper body actions and the whole body actions (jogging, walking, and running.)

TABLE IV
CONFUSION MATRIX OF ACTION RECOGNITION ON MSR ACTION DATASET II (OUR METHOD)

	boxing	clapping	waving
boxing	**61**	8	12
clapping	2	**38**	11
waving	10	0	**61**

TABLE V
CONFUSION MATRIX OF ACTION RECOGNITION ON MSR ACTION DATASET II (METHOD OF YUAN *ET AL.* [55])

	boxing	clapping	waving
boxing	**40**	0	41
clapping	2	**21**	28
waving	11	0	**60**

We further compare our results to the state-of-the-art results in [55] and validate the effectiveness of the proposed hierarchical filtered motion field. Yuan *et al.* [55] applied exactly the same STIP feature extraction and HOG/HOF descriptors as Laptev *et al.* [22] but a naïve-Bayes based mutual information maximization (NBMIM) for action classification. For method of Yuan *et al.* [55], the NBMIM classifier is trained on KTH dataset for 4 action classes (boxing, hand clapping, hand waving, and others) and tested on MSR Action Dataset II for 3 classes: boxing, hand clapping, and hand waving. The action recognition rate on MSR Action Dataset II is 59.6%. The confusion matrix is displayed in Table V.

TABLE VI
COMPARISON WITH THE STATE-OF-THE ART RESULTS ON MSR ACTION DATASET II (3 ACTIONS WITH MOVING BACKGROUND) BY USING THE CLASSIFIER TRAINED ON 6 CLASSES FOR GMM AND 4 CLASSES FOR NBMIM ON KTH DATASET

Method	Accuracy
2D corners + IP Detection + Hierarchical motion filter + HOG/HOG-MHI + GMM	**78.8%**
2D corners + IP Detection + HOG/HOG-MHI + GMM	**71.4%**
STIP + HOG/HOF + NBMIM (Yuan *et al.* [55])	59.6%

To validate the effectiveness of the proposed hierarchical filtered motion field in handling crowded videos, we compare the results of with/without using it. As we described in previous sections, we extract the 2D corners with recent motion as interest points and represent the features as HOG/HOG-MHI at each interest point. As shown in Table VI, we achieve 71.4% recognition rate on MSR Action Dataset II without using the hierarchical filtered motion field. The recognition rate is improved to 78.8% with using the hierarchical motion filter.

D. *Cross-dataset Action Detection Results*

We further perform cross-dataset test for action detection, which not only detect the category of actions, but also the spatio-temporal locations of action instances in a video sequence. For action detection, we use a 3D subvolume C to represent a cuboid in the 3D video space [*x1, x2, y1, y2, t1, t2*] that contains an action instance. The spatial locations of the subvolume [*x1, x2, y1, y2*] identify where the action happens, while the temporal locations of the subvolume [*t1, t2*] denote when the action happens. In order to find the optimal subvolumes C^* containing the action of interests for a given video sequence, the score of each interest point is computed by employing background model as following equation [5]:

$$f(x) = \ log \frac{\Pr(x|\Omega^c)\Pr(\Omega^c)}{\Pr(x|\Omega^b)\Pr(\Omega^b)} \qquad (8)$$

Where Ω^c and $\Omega^b$are the GMM models for the interested action and the background with the corresponding prior distributions $\Pr(\Omega^c)$ and $\Pr(\Omega^b)$ respectively. The best subvolumes C^* can be represented as:

$$\mathrm{C*}{=}\,\mathrm{argmax}_C \sum_{x\epsilon V} f(x) \qquad (9)$$

For each action type, after the score of each interest point is computed, a spatio-temporal branch-and-bound algorithm [55]55] is used to find all the subvolumes whose total scores are above a threshold (By varying the threshold, we obtain the precision-recall curve as we will show later). The spatio-temporal branch-and-bound code is downloaded from their website [60][60].

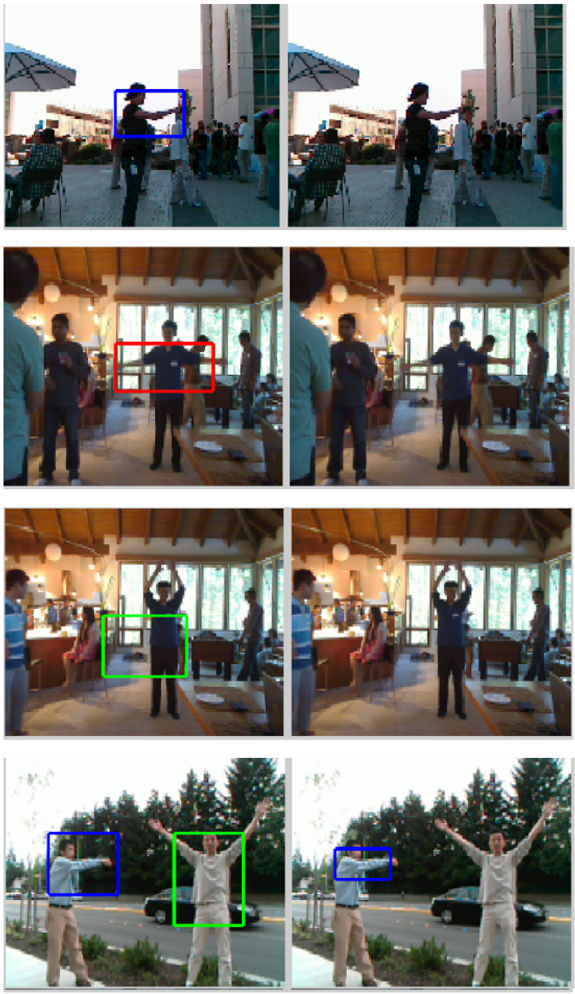
Again, we use KTH dataset for training and MSR Action Dataset II for testing. We find that the detection results of using Hierarchical Filtered Motion features are significantly better than those of using STIP features [22]. Figure 9 shows some examples when STIP feature fails while Hierarchical Filtered Motion feature works well.

To provide a quantitative evaluation, we measure the precision and recall of our detection results. The precision and recall are defined as:

$$Recall = \frac{number\ of\ detected\ groundtruth}{number\ of\ groundtruth}$$

$$Precision = \frac{number\ of\ correctly\ detected\ bounding\ boxes}{number\ of\ detected\ bounding\ boxes}$$

We vary the detection threshold to obtain the precision-recall curves one for each action type, as shown in Figure 10. It can be seen that our Hierarchical Filtered Motion feature is significantly better than STIP feature for all three action types.

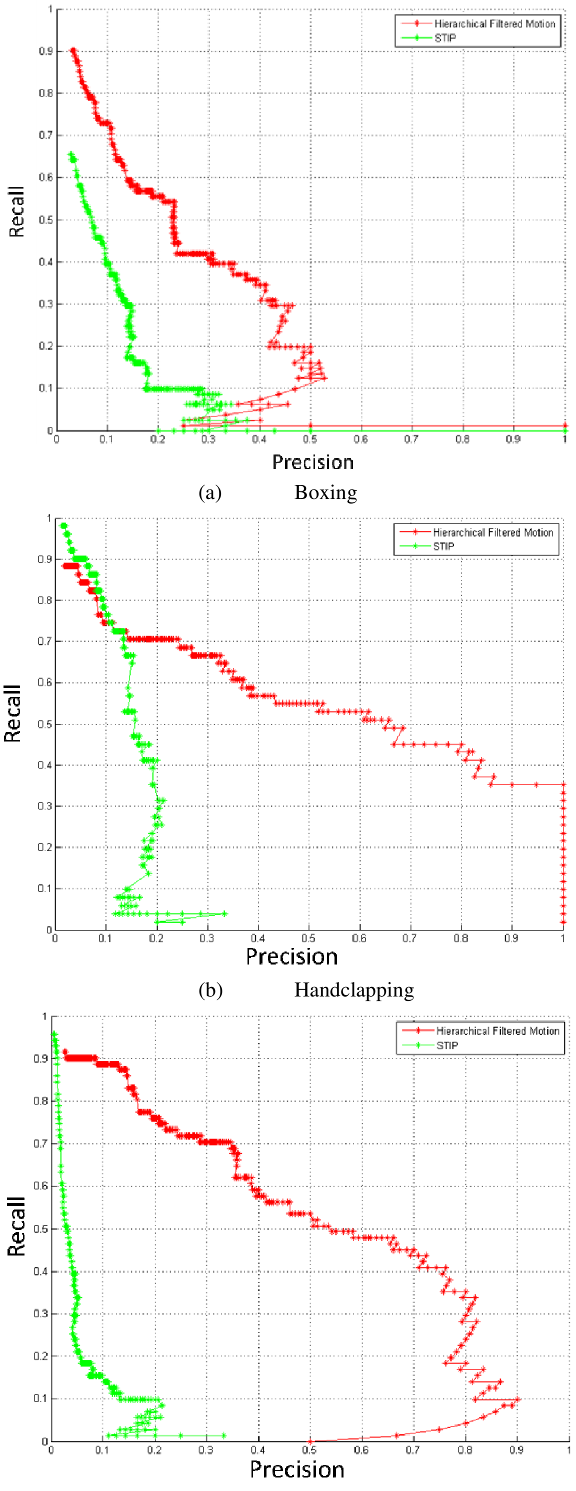


(a) Hierarchical Filtered Motion (b) STIP

Fig. 9. Examples where Hierarchical Filtered Motion features successfully detect the action of interests while STIP features fail. For each example, the left picture illustrates the detection results of using Hierarchical Filtered Motion features, while the right shows the results of using STIP features. The three colors denote different kinds of actions: red for clapping, green for waving, and blue for boxing.

E. *Algorithm Efficiency Analysis*

The proposed Hierarchical Filtered Motion feature extraction is programmed in C++ without optimization. The speed of Hierarchical Filtered Motion feature extraction in images at resolution of 160x120 of MSR dataset is approximately 46 frames per second. In comparison, STIP feature extraction is 10 frames per second. The testing is performed on a computer with Duo CPU (2.2GHz and 1.18GHz) with 3.49GB memory for both algorithms which include loading video, displaying features and saving the extracted features to a file.



(a) Boxing

(b) Handclapping

(c) Hand waving

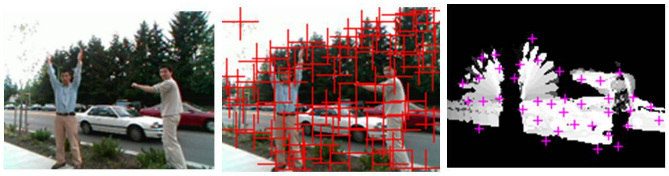
Fig. 10. The precision-recall curves of action detection using Hierarchical Filtered Motion (red color) and STIP (green color) features. (a) boxing; (b) hand clapping; (c) hand waving.

We further extensively investigate the efficiency of the Hierarchical Filtered Motion feature extraction of the following steps: 1) interest point detection including 2D Harris Corner Detection, MHI calculation, and removing the corners in the static background using MHI as the motion mask; 2) hierarchical motion filter feature extraction including processing of both global and local motion filters and calculation of HOG and MHI-HOG; and 3) the total computation time of step 1 and 2. We use MSR dataset with crowded background motions. One example of the testing data is displayed in Figure 11. The average number of detected interest points is about 30 per image for the tested sequence. The details of the efficiency of the proposed Hierarchical Filtered Motion feature extraction are listed in Table VII. For the sequence with resolution of 160x120, the speed of interest point detection (step 1) is 216 frames per second. The speed for Hierarchical Filtered Motion feature extraction (step 2) is 98 frames per second. The speech of the whole core algorithm (step 1 + step 2) is 68 frames per second (without loading video, displaying features and saving the extracted features to a file). The above speeds decrease to 90, 45, and 30 frames per second for sequence in resolution of 320x240. Note that to keep the same amount of Harris corners in both resolutions, we double the minimum distance between corners in 2D Harris corner detection for 320x240 images.

TABLE VII

EFFICIENCY ANALYSIS FOR HIERARCHICAL FILTERED MOTION FEATURE EXTRACTION ON MSR DATASET WITH CROWDED BACKGROUND

Image resolution (MSR dataset)	**Efficiency (frame/second)**		
	IP detection	Hierarchical motion filter feature extraction	Total
160x120	216	98	68
320x240	90	45	30



(a) Original image (b) 2D Harris corners (c) MHI filtered corners

Fig. 11. Example images from MSR dataset for efficiency analysis.

V. CONCLUSION

We have presented a new feature for action recognition in crowded videos without tracking objects or key points. A novel technique, called Hierarchical Filtered Motion, was proposed to reduce distracting motions caused by the background moving objects near an interest point. We have performed action classification and detection experiments on videos with cluttered and moving background, and demonstrated its superior performance over existing techniques. In addition, our approach is very fast thus suitable for real-time action recognition.

The proposed Hierarchical Filtered Motion is more robust than STIP features for action recognition in crowded videos. The reasons are summarized as the following: 1) the 2D Harris corner detection is less sensitive to lighting changes than STIP features; 2) MHI filtered interest points can better characterize the motion features than STIP (too sparse); 3) The directional motion information is measured directly from the intensity gradients in the MHI. It is also more robust because the motion information in MHI is mainly along the contours of the moving objects. Thus, unwanted motion in the interior regions of object contours is ignored; and 4) the Hierarchical Filtered Motion computes a structure proximity between any pixel in the local region and the interest point and can reduce distracting motions caused by the background moving objects near an interest point. Our future work will focus on recognizing more types of actions in crowded videos.

Acknowledgments

The authors thank the anonymous reviewers for their constructive comments and suggestions that help to improve the quality of this manuscript.

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] A.F. Bobick and J.W. Davis, The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. 23, 257–267, 2001.

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

[6] L. Cao, Y. Tian, Z. Liu, B. Yao, Z. Zhang, and T. Huang, Action Detection using Multiple Spatial-Temporal Interest Point Features, IEEE International Conference on Multimedia & Expo (ICME), 2010.

[7] N. Dalal and B. Triggs, Histograms of Oriented Gradients for Human Detection, CVPR, 2005.

[8] J.W. Davis, "Hierarchical motion history images for recognizing human motion", Proc. Of IEEE Workshop on. Detection and Recognition of Events in Video, 2001.

[9] P. Dollar, V. Rabaud, G. Cottrell, S. Belongie,: Behavior recognition via sparse spatio-temporal features. In: Proc. of ICCV Int. work-shop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VSPETS), pp. 65–72, 2005.

[10] A. Fathi and G. Mori. Action recognition by learning mid-level motion features. *CVPR*, 2008.

[11] 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.

[12] A. Gilbert, J. Illingworth, and R. Bowden. Scale invariant action recognition using compound features mined from dense spatio-temporal corners. *ECCV*, pages 222-233, 2008.

[13] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. Alvey Vision Conf.*, pages 189–192, 1988.

[14] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. ICCV, 2007.

[15] X. Ji and H. Liu, Advances in View-Invariant Human Motion Analysis: A Review, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, Volume: 40, Issue: 1, p13-24, 2010.

[16] I. Junejo, E. Dexter, I. Laptev, and P Perez, View-Independent Action Recognition from Temporal Self-Similarities**,** IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume: 33, Issue: 1 2011 , Page(s): 172 – 185.

[17] M. B. Kaaniche and F. Bremond, Gesture Recognition by Learning Local Motion Signatures, In CVPR, 2010.

[18] Y. Ke, R. Sukthankar, and M. Hebert, Event Detection in Crowded Videos, ICCV, 2007.

[19] A. Kovashka and K. Grauman, Learning a Hierarchy of Discriminative Space-Time Neighborhood Features for Human Action Recognition, In CVPR, 2010.

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

[21] I. Laptev and T. Lindeberg. Space-time interest points. In ICCV, 2003.

[22] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning Realistic Human Actions from Movies, CVPR, 2008.

[23] G. Lavee, E. Rivlin, and M. Rudzsky, Understanding Video Events: A Survey of Methods for Automatic Interpretation of Semantic Occurrences in Video, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, Volume: 39 , Issue: 5, p489-504, 2009.

[24] N. Li, C. Chen, Q. Wang, M. Song, D. Tao, and X. Li, "Avatar Motion Control by Natural Body Movement via Camera," Neurocomputing (Elsevier), vol. 72, no. 1-3, pp. 648-652, 2008.

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

[26] J. Liu and M. Shah. Learning human actions via information maximization. *CVPR*, 2008.

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

[28] M. Marszałek, I. Laptev, and C. Schmid. Actions in context. In CVPR, 2009.

[29] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, Nov 2004.

[30] R. Mattivi and L. Shao, Human Action Recognition Using LBP-TOP as Sparse Spatio-Temporal Feature Descriptor, Computer Analysis of Images and Patterns, Springer, 2009.

[31] H. Meng, N. Pears, and C. Bailey, Motion Feature Combination for Human Action Recognition in Video, J. Braz et al. (Eds.): VISIGRAPP 2007, CCIS 21, 2008.

[32] R. Messing, C. Pal, and H. Kauze, Activity Recognition Using the Velocity Histories of Tracked Keypoints, ICCV, 2009.

[33] K. Mikolajczyk and H. Uemura. Action recognition with motion-appearance vocabulary forest. *CVPR*, 2008.

[34] M. Nicolescu and G. Medioni, "A Voting-Based Computational Framework for Visual Motion Analysis and Interpretation", *IEEE Trans. on PAMI*, vol. 27, no. 5, 2005.

[35] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *IJCV,* 79(3):299-318, 2008.

[36] A. Oikonomopoulos, I. Patras, and M. Pantic. Spatio-temporal salient points for visual recognition of human actions. IEEE Trans. Systems, Man, and Cybernetics, Part B, 36, (3):710–719, 2006.

[37] Y. Pang, X. Li, Y. Yuan, D. Tao, and J. Pan, Fast Haar Transform based Feature Extraction for Face Representation and Recognition, IEEE Transactions on Information Forensics & Security (T-IFS), vol. 4, no. 3, pp. 441-450, 2009.

[38] Y. Pang, Y. Yuan, and X. Li, Gabor-based Region Covariance Matrices for Face Recognition, IEEE Transactions on Circuits and Systems for Video Technology (T-CSVT), vol. 18, no. 7, pp. 989-993, 2008.

[39] Y. Pang, Y. Yuan, and X. Li, Effective Feature Extraction in High-Dimensional Space. IEEE Transactions on Systems, Man, and Cybernetics, Part B 38(6): 1652-1656, 2008.

[40] Y. Pang, Y. Yuan, X. Li, and J. Pan, Efficient HOG human detection, Signal Processing, Vol. 91, Issue 4, Pages 773-781, 2011.

[41] D. A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, Speaker Verification Using Adapted Gaussian Mixture Models, Digital Signal Processing, vol. 10, pp. 19–41, 2000.

[42] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *ICPR*, pages III: 32–36, 2004.

[43] J. Shen, D. Tao, and X Li, Modality Mixture Projections for Semantic Video Event Detection. IEEE Trans. Circuits Syst. Video Techn. 18(11): 1587-1596, 2008.

[44] J. Sun, X. Wu, S. Yan, L. Cheong. T. Chua, and J. Li, Hierarchical Spatio-Temporal Context Modeling for Action Recognition, CVPR, 2009.

[45] Y. Tian, R. Feris, H. Liu, A. Hampapur, and M. Sun, Robust Detection of Abandoned and Removed Objects in Complex Surveillance Videos, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, Volume: PP , Issue: 99, p1-12, 2010.

[46] H. Wang, M. Ullah, A. Klaser, I. Laptev, and C. Schmid, Evaluation of local spatio-temporal features for action recognition, BMVC, 2009.

[47] 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.

[48] Y. Wang and G. Mori, Hidden Part Models for Human Action Recognition: Probabilistic vs. Max-Margin, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume: PP, Issue: 99, 2010.

[49] D. Weinland, R. Ronfard, and E. Boyer, Free viewpoint action recognition using motion history volumes, Computer Vision and Image Understanding, Volume 104 , Issue 2, November 2006.

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

[51] 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.

[52] S Yan, X Zhou, M Liu, M Hasegawa-Johnson, and T. Huang, Regression from patch-kernel, CVPR 2008.

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

[54] J. Yin and Y. Meng, Human Activity Recognition in Video using a Hierarchical Probabilistic Latent Model, In CVPR, 2010.

[55] J. Yuan, Z. Liu, and Y. Wu, Discriminative Subvolume Search for Efficient Action Detection, CVPR, 2009.

[56] Y. Yuan, Y. Pang, J. Pan, and X. Li, Scene Segmentation Based on IPCA for Visual Surveillance, Neurocomputing (Elsevier), vol. 72, nos. 10-12, pp. 2450-2454, 2009.

[57] Y. Yuan and Y. Pang, “Discriminant Adaptive Edge Weights for Graph Embedding,” IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 1993-1996, 2008.

[58] H. Zhou, Y. Yuan, et al., “Non-rigid Object Tracking in Complex Scenes,” Pattern Recognition Letters (Elsevier), vol. 30, no. 2, pp. 98-102, 2009.

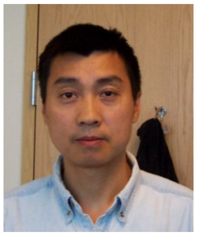
[59] G. Zhu, M. Yang, K. Yu, W. Xu, and Y. Gong. Detecting video events based on action recognition in complex scenes using spatio-temporal descriptor. In *Proc. ACM international conference on Multimedia*, pages 165–174, 2009.

[60] MSR Action Dataset II, http://research.microsoft.com/en-us/um/people/zliu/ActionRecoRsrc/.

YingLi Tian (M’99–SM’01) received her BS and MS from TianJin University, China in 1987 and 1990 and her PhD from the Chinese University of Hong Kong, Hong Kong, in 1996. After holding a faculty position at National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, she joined Carnegie Mellon University in 1998, where she was a postdoctoral fellow of the Robotics Institute. Then she worked as a research staff member in IBM T. J. Watson Research Center from 2001 to 2008. She is one of the inventors of the IBM Smart Surveillance Solutions.

She is currently an associate professor in Department of Electrical Engineering at the City College of New York. Her current research focuses on a wide range of computer vision problems from activity detection and analysis, to human identification, facial expression analysis, and video surveillance. She is a senior member of IEEE.

Liangliang Cao is a Ph.D. student in the Department of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign (UIUC). He received my Bachelor of Engineering on Electronic Engineering from University of Science and Technology of China (USTC) and Master of Philosophy on Information Engineering from The Chinese University of Hong Kong (CUHK). He received the USTC Outstanding Undergraduate Research Award in 2003, and was nominated for the CUHK Distinguished M. Phil. Thesis of the Faculty of Engineering in 2005. He was awarded as a UIUC Computational Science and Engineering Fellow in 2009, a Facebook Fellowship Finalist and an “Emerging Leader in Multimedia and Signal Processing” in IBM Watson workshop in 2010. He and his colleagues participated in TRECVID 2008 Airport Surveillance Competition and ImageNet 2010 Large Scale Visual Recognition Challenge and was ranked 1st place in both competitions. His research interests include web-scale multimedia, computer vision, data mining and distributed computing.



Zicheng Liu (SM'05) received the B.S. degree in mathematics from Huazhong Normal University, Wuhan, China, M.S. degree in operational research from the Institute of Applied Mathematics, Chinese Academy of Sciences, Beijing, and the Ph.D. degree in computer science from Princeton University, Princeton, NJ.

He is a senior researcher at Microsoft Research, Redmond, WA. He has worked on a variety of topics including combinatorial optimization, linked figure animation, and microphone array signal processing. His current research interests include activity recognition, face modeling and animation, and multimedia collaboration. Before joining Microsoft Research, he worked at Silicon Graphics as a member of technical staff for two years where he developed a trimmed NURBS tessellator which was shipped in both OpenGL and OpenGL-Optimizer products. He has published over 70 papers in peer-reviewed international journals and conferences, and holds over 40 granted patents. He has coauthored a book entitled Face Geometry and Appearance Modeling (Cambridge, U.K., Cambridge Univ. Press, 2011).

Dr. Liu has served in the technical committees for many international conferences. He was the co-chair of the 2003 ICCV Workshop on Multimedia Technologies in E-Learning and Collaboration, the technical co-chair of 2006 IEEE International Workshop on Multimedia Signal Processing, and the technical co-chair of 2010 International Conference on Multimedia and Expo. He is an associate editor of Machine Vision and Applications journal.

Zhengyou Zhang received the B.S. degree in electronic engineering from Zhejiang University, Hangzhou, China, in 1985, the M.S. degree in computer science from the University of Nancy, Nancy, France, in 1987, and the Ph.D. degree in computer science and the Doctorate of Science (*Habilitation à diriger des recherches*) from the University of Paris XI, Paris, France, in 1990 and 1994, respectively.

He is a Principal Researcher with Microsoft Research, Redmond, WA, USA, and manages the multimodal collaboration research team. Before joining Microsoft Research in March 1998, he was with INRIA (French National Institute for Research in Computer Science and Control), France, for 11 years and was a Senior Research Scientist from 1991. In 1996-1997, he spent a one-year sabbatical as an Invited Researcher with the Advanced Telecommunications Research Institute International (ATR), Kyoto, Japan. He has published over 200 papers in refereed international journals and conferences, and has coauthored the following books: *3-D Dynamic Scene Analysis: A Stereo Based Approach* (Springer-Verlag, 1992); *Epipolar Geometry in Stereo, Motion and Object Recognition* (Kluwer, 1996); *Computer Vision* (Chinese Academy of Sciences, 1998, 2003, in Chinese); *Face Detection and Adaptation* (Morgan and Claypool, 2010), and *Face Geometry and Appearance Modeling* (Cambridge University Press, 2011). He has given a number of keynotes in international conferences.

Dr. Zhang is a Fellow of the *Institute of Electrical and Electronic Engineers* (IEEE), the Founding Editor-in-Chief of the *IEEE Transactions on Autonomous Mental Development*, an Associate Editor of the *International Journal of Computer Vision*, and an Associate Editor of *Machine Vision and Applications*. He served as Associate Editor of the *IEEE Transactions on Pattern Analysis and Machine Intelligence* from 2000 to 2004, an Associate Editor of the *IEEE Transactions on Multimedia* from 2004 to 2009, among others. He has been on the program committees for numerous international conferences in the areas of autonomous mental development, computer vision, signal processing, multimedia, and human-computer interaction. He served as a Program Co-Chair of the *International Conference on Multimedia and Expo* (ICME), July 2010, a Program Co-Chair of the ACM *International Conference on Multimedia* (ACM MM), October 2010, and a Program Co-Chair of the ACM *International Conference on Multimodal Interfaces* (ICMI), November 2010. He is serving a General Co-Chair of the IEEE *International Workshop on Multimedia Signal Processing* (MMSP), October 2011.