# Action Detection by Fusing Hierarchically Filtered Motion With Spatiotemporal Interest Point Features

**YingLi Tian**
*The City College, City University of New York, New York, NY 10031, USA*

**Liangliang Cao,**
*IBM T. J. Watson Research Center, Hawthorne, NY 10532, USA*

**Zicheng Liu**
*Microsoft Research, Redmond WA 98052 USA*

**Zhengyou Zhang**
*Microsoft Research, Redmond WA 98052 USA*

## ABSTRACT

This chapter addresses the problem of action detection from cluttered videos. In recent years, many feature extraction schemes have been designed to describe various aspects of actions. However, due to the difficulty of action detection, *e.g.*, the cluttered background and potential occlusions, a single type of features cannot effectively solve the action detection problems in cluttered videos. In this chapter, we propose a new type of features, Hierarchically Filtered Motion (HFM), and further investigate the fusion of HFM with Spatiotemporal Interest Point (STIP) features for action detection from cluttered videos. In order to effectively and efficiently detect actions, we propose a new approach which combines Gaussian Mixture Models (GMMs) with Branch-and-Bound search to locate interested actions in cluttered videos. The proposed new HFM features and action detection method have been evaluated on the classical KTH dataset and the challenging MSR Action Dataset II which consists of crowded videos with moving people or vehicles in background. Experiment results demonstrate that the proposed method significantly outperforms existing techniques especially for action detection in crowded videos.

## INTRODUCTION

In the past few years, computer vision researchers have witnessed a surge of interest in human action analysis through videos. Human action recognition, *which classifies a video to a predefined action category*, was first studied under well controlled laboratory scenarios, e.g., with clean background and no occlusions (Schuldt *et al*., 2004). Later research work shows that action recognition is important for analyzing and organizing online videos (Liu *et al*., 2009). Moreover, action recognition plays a crucial role in building surveillance system (Hu *et al*., 2009) and studying customer behaviors. With the increasing of web video clips (*e.g.*, videos on Youtube) and the surveillance systems, it has become very important to effectively analyze video actions.

An effective analysis of video actions requires action detection, *which can not only answer "which action happens in a video", but also "when and where the action happens in the video sequence"*. In other words, action detection will detect action category, locations, and time in video sequences than simply classifying a video clip to one of the existing action labels. When a video contains multiple actions, simple action classification will not work. In practice, surveillance videos often contain multiple types of actions, where only action detection can provide meaningful results.

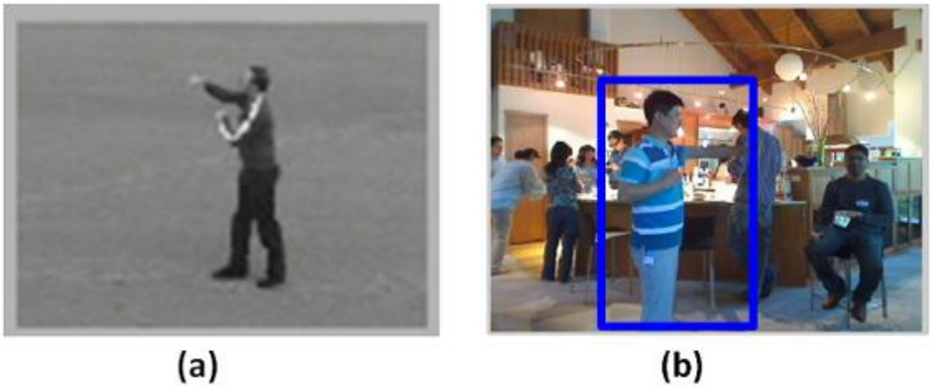


*Figure 1. Comparing the differences between action classification and detection. (a) For a classification task we need only estimate the category label for a given video. (b) For an action detection task we need not only estimate the category of an action but also the location of the action instance.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

Action detection is a challenging task. As shown in Figure 1, the background is often cluttered, and the crowds might occlude each other in complex scenes. It is difficult to distinguish interested actions from other video contents. The appearance of interested actions might have similar appearance of the background. Furthermore, the motion field of an action might be occluded by other moving objects in the scene. Due to the difficulty of locating human actions, most existing datasets of human actions (Blank *et al*., 2005, Schuldt *et al*., 2004) only involve action classification task without detecting locations of actions, where human actions are usually recorded with clean backgrounds, and each video clip mostly involves a single person who repeatedly performs one category of actions within a whole video clip.

In this chapter, we address the action detection problem by proposing a new type of features, Hierarchically Filtered Motion (HFM), and further investigate the fusion of HFM with other Spatiotemporal Interest Point (STIP) features (Dollar *et al*., 2005, Laptev and Lindeberg, 2003, Cao *et al*. 2010, Tian *et al*., 2011) for action detection from cluttered videos. An action is often associated with multiple visual measurements, which can be either appearance features (e.g., color, edge histogram) or motion features (e.g., optical flow, motion history). Different features describe different aspects of the visual characteristics and demand different metrics. How to handle heterogeneous features for action detection becomes an important problem.

The difficulty of combining multiple features lies in the heterogeneous nature of different features. Different STIP features are based on different detectors, and the number of detected features varies significantly. It is still an open question how to effectively combine such features. A naive approach is to quantize STIP features and build histogram based on quantization indices. However, much information will be lost in the quantization process, and a histogram representation overlooks the differences in the number of detected features. Therefore, simply combining histograms will produce poor detection results.

Our work employs a probabilistic representation of the different features so that we can quantitatively evaluate the contribution from each of these features. We estimate the likelihood of each feature vector belonging to a given action of interests, which can be viewed as normalized contribution from different features. The optimal bounding box corresponds to the maximum likelihood and is found by a branch-bound search. In our approach, we model each feature vector with Gaussian Mixture Models (GMMs). GMMs with large number of components are known to have the ability to model any given probability distribution function. Based on GMMs, we can estimate the likelihood of each feature vector belongings to a given action of interests. The likelihood can be viewed as a normalized contribution from different features, and the optimal bounding box corresponds to the maximum likelihood. The bounding box is found by a branch-bound search (Yuan *et al*., 2009), which is shown to be efficient and effective to locate the action of interest.

## BACKGROUND

Many approaches have been proposed on action recognition (Jhuang *et al*., 2007, Laptev *et al*., 2008, Liu *et al*., 2009, Messing *et al*., 2009, Reynolds *et al*., 2000, Yuan *et al*., 2009). Compared with the task of action classification, action detection is more challenging. There are only a few literatures devoted to the task of action detection (Du and Yuan, 2011, Cao *et al*., 2009, Cao *et al*., 2010, Hu *et al*., 2009, Ke *et al*., 2007, Yao and Zhu, 2009, Yuan and Pang, 2008). Laptev *et al.* (Laptev *et al*., 2008, Marszałek *et al*., 2009) used local spatiotemporal invariant points (STIPs) (Laptev and Lindeberg 2003), space-time pyramids, local spatiotemporal descriptors (HOG/HOF) (Dalal and Triggs, 2005, Pang *et al*., 2010), and multichannel non-linear SVMs for realistic actions in movies. Yuan *et al.* (Yuan *et al*., 2009) 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.* (Ke *et al*., 2007) proposed a method to detect event in crowded videos by combining spatiotemporal shapes with a flow descriptor. Sun *et al.* (Sun *et al*., 2009) modeled the spatiotemporal 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.* (Messing *et al*., 2009) 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.* (Junejo *et al*., 2011) attempted to recognize human actions under different views using temporal self-similarities. Yin and Meng (Yin and Meng, 2010) 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 paper (Ji and Liu, 2010).

Despite promising results are achieved by the state-of-the-art work, 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 STIP detectors and descriptors, and 2) the trajectories of keypoints cannot be reliably tracked in crowded videos. A majority of recent work on STIPs takes quantized STIPs as input and builds histograms based on quantization indices. The quantized STIPs are also called video codewords. The histograms can be fed into discriminative SVM classifiers (Schuldt *et al*., 2004) or generative topic models (Niebles *et al*., 2006). The collection of quantized codewords is also named as a codebook. The use of the codebook and histogram is preferred because it can condense different number of STIPs into a fixed length feature vector. However, quantized codeword representation is not a good fit for cross-dataset scenarios due to the large variations of STIPs in different environments. Given two videos captured with different viewing points and light conditions, the corresponding distributions of STIPs are likely quite different. If we build a new codebook on a new dataset, the word histogram representation will be totally different, so the old model cannot be directly applied the new dataset. In summary, quantized codeword representation overlooks the differences of STIP distributions in different environments and may fail to correctly transfer the knowledge from source dataset to target dataset.

Motivated by the recent success of SIFT and HOG in image domain, many researchers have designed various counterparts to describe the spatial salient patches in video domain. Laptev and Lindeberg (Laptev and Lindeberg, 2003) generalized Harris detector to spatiotemporal space. They aim to detect image patches with significant local variations in both space and time and compute their scale invariant spatiotemporal descriptors. This approach is later improved by (Laptev *et al.*, 2008) which gives up scale selection but uses a multi-scale approach and extract features at multiple levels of spatiotemporal scales. The improved method yields reduced computational complexity, denser sampling, and suffers less from scale selection artifacts. Another important video feature is designed by Dollar *et al.* (Dollar *et al.*, 2005), which detects the salient patches by finding the maximum of temporal Gabor filter responses. This method aims to detect regions with spatially distinguishing characteristics undergoing a complex motion. In contrast, patches undergoing pure translational motion, or patches without spatially distinguishing features will in general not induce a response. After the salient patches are detected, the histogram of 3D cuboid is introduced to describe the patch feature.

Many action classification systems (Dollar *et al.*, 2005, Jhuang *et al.*, 2007, Laptev and Lindeberg, 2003, Niebles *et al.*, 2006, Wong *et al.*, 2007, Wu *et al.*, 2007) are built using Laptev's or Dollar's features. These two features focus short-term motion information instead of long term motion, and motion field of a salient patch sometime is contaminated by the background motions. However, most of existing systems only classify video clips to predefined action categories, and does not consider the location task.

To overcome the limitations of existing salient patch descriptors, a hierarchically filtered motion field method has been proposed recently for action recognition (Tian *et al.*, 2011). This work applies global spatial motion smoothing filter to eliminate isolated unreliable or noisy motions. To characterize long-term motion features, Motion History Image (MHI) is employed as basic representations of interest points. This new feature is named as Hierarchically Filtered Motion (HFM) and works well in crowded scenes. We believe the HFM describes complementary aspects of video actions and this work will combine HFM with the existing features of (Dollar *et al.*, 2005, Laptev *et al.*, 2008) for action detection tasks.

The existing work of action detection (Du *et al.*, 2011, Cao *et al.*, 2009, Cao *et al.*, 2010, Hu *et al.*, 2009, Ke *et al.*, 2007, Yao *et al.*, 2009, Yuan *et al.*, 2008) only use single type of features. Although multiple feature fusion was proved to be effective in action classification (Cao *et al.*, 2009, Liu *et al.*, 2008), it is still an untouched problem to combine multiple features for action detection.

The difficulty of applying multiple features for action detection is two-fold: First, existing fusion methods (Cao *et al.*, 2009, Liu *et al.*, 2008] assume that each sample has the same number of features. However, in action detection, different features correspond to different detectors, and the numbers of detected salient patches are usually different subject to different features. Second, detecting actions in videos involves a searching process in x-y-t dimensions, which is very computationally expensive. Many existing feature fusion methods (Cao *et al.*, 2009) are usually too slow for this task. This chapter employs Gaussian Mixture Models (GMMs) to model heterogeneous features, and the probability of a given feature vector is estimated effectively based on the GMM model. To locate the action of interests, we employ a branch-and-bound method to find the optimal subvolumes which correspond to the largest GMM scores. Note that although this chapter only combines three types of features from (Dollar *et al.*, 2005, Laptev *et al.*, 2008, Tian *et al.*, 2011), our method is a general framework and can be used to fuse more features (Boiman *et al.*, 2005, Rodriguez *et al.*, 2008, Zhu *et al.*, 2009).

## FUSION OF MULTIPLE SPATIOTEMPORAL INTEREST POINT FEATURES

In this section, we describe the features we used for action detection from cluttered videos by fusion of multiple spatiotemporal interest point features including: Hierarchically Filtered Motion (HFM) (Tian *et al.,* 2011), and other Spatiotemporal Interest Point (STIP) features (Dollar *et al.*, 2005, Laptev and Lindeberg, 2003).

### 1. Hierarchically Filtered Motion Features

Hierarchically Filtered Motion (HFM) features employ Motion History Image (MHI) (Bobick *et al.*, 2001, Davis, 2001] as basic representations of motion due to its robustness and efficiency. First, we detect interest points as 2D Harris corners with recent motion, e.g. locations with high intensities in MHI 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. The spatial and temporal features are then described by Histograms of Oriented Gradient (HOG) in the intensity image and MHI respectively.

**Motion History Image (MHI):** MHI is a real-time motion template that temporally layers consecutive image differences into a static image template (Bobick and Davis, 2001, Davis, 2001). 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 (Bobick and Davis, 2001). 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 $\tau$ is the maximum duration of motion. We set $\tau$ 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.

In order to handle cluttered background, we propose a hierarchically filtered motion field technique based on Motion Gradient Image (MGI). The MGI is the intensity gradients of MHI which directly yield motion orientations. Note that the magnitudes of 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.

**Hierarchically Filtered Motion Features:** To remove the isolated distracting motions, hierarchically filtered motion features includes two levels of processing: a *global* spatial motion smoothing filter to the gradients of MHI and a *local* motion field filter by computing a structure proximity between any pixel in the local region and the interest point. 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. 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. In our implementation, we use an 8-bin-layer representation of a binary image of the smoothed gradients of MHI. At each interest point, a local filtered motion field is applied 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. 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 should be enhanced. On the other hand, the blobs with longer distances to the interest point most likely belong to other objects. 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. The structure proximity values are used to normalize motion histograms in HOG-MHI calculation. More details can be found in paper (Tian *et al*., 2011).

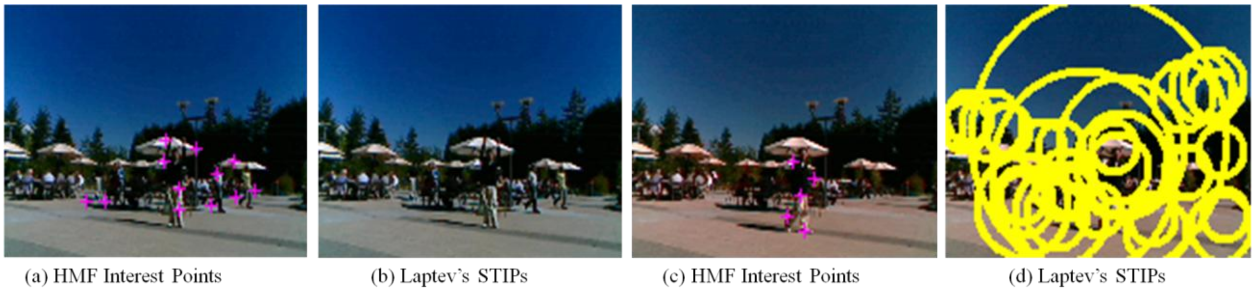
## 2. Fusion of Multiple Spatiotemporal Interest Points Features

Sparse selection of STIPs has been successfully used for action recognition (Cao *et al*., 2010, Dollar *et al*., 2005, Laptev and Lindeberg, 2003, Laptev *et al*., 2008, Mattivi and Shao, 2009, Niebles *et al*., 2006, Schuldt *et al*., 2004, Yuan *et al*., 2009). Laptev *et al.* developed a nice mathematic framework to find pixels with significant variations in both spatial and temporal directions (Laptev and Lindeberg, 2003). 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 (Dollar *et al*., 2005). The limitation of the approach in (Dollar *et al*., 2005) 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.

**Laptev's STIP Features:** Laptev and Lindeberg (Laptev and Lindeberg, 2003) generalized Harris detector to spatiotemporal space. They aim to detect image patches with significant local variations in both spatial and temporal directions, and compute the scale-invariant spatiotemporal descriptors. This approach is later improved by (Laptev *et al*., 2008) which gives up scale selection but uses a multi-scale approach and extract features at multiple levels of spatiotemporal scales. The improved method yields reduced computational complexity denser sampling, and suffers less from scale selection artifacts.

**Dollar's STIP Features:** Another important type of video features is designed by Dollar *et al.* (Dollar *et al*., 2005), which detects the salient patches by finding the maximum of temporal Gabor filter responses. This method aims to detect regions with spatially distinguishing characteristics undergoing a complex motion. In contrast, patches undergoing pure translational motion, or patches without spatially distinguishing features will in general not induce a response. After the salient patches are detected, the histogram of 3D cuboid is introduced to describe the patch feature.

In this chapter, we investigate the combination multiple STIPs (Laptev *et al*., 2008) for the videos with cluttered background. We observe that there are not enough interest points in action regions for some video sequences (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 (Harris and Stephens 1988) 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.


(a) HMF Interest Points

(b) Laptev's STIPs

(c) HMF Interest Points

(d) Laptev's STIPs

*Figure 2: Examples of interest point detection by our method and STIP detection of Laptev et al. (Laptev* and Lindeberg *2003, Laptev et al., 2008) 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 (Laptev* and Lindeberg *2003, Laptev et al., 2008); (c) our interest point detection is insensitive to lighting changes; and (d) false STIPs are detected on background regions (Laptev* and Lindeberg *2003, Laptev et al., 2008).*

## 3. HOG-based Feature Descriptor and Representation for Action Detection

**HOG and HOF Feature Descriptor for Latpev's STIP Features:** Histograms of Oriented Gradients (HOG) feature descriptors have been widely used in human detection (Dalal and Triggs, 2005, Laptev *et al*., 2008, Marszałek *et al*., 2009, Pang *et al*., 2010, Yuan *et al*., 2009). Laptev *et al*. (Laptev *et al*., 2008) employ HOG and Histograms of Optical Flow (HOF) descriptors for 3D video patches in the neighborhood of detected STIPs. However, these features are based on single frame or neighboring frames, but overlooks the motion descriptions over a longer time. For each interest point, the local appearance and motion are characterized by grids of HOG and HOF respectively in the neighborhood with a window size. For each grid, 8 directions are employed for both HOG and HOF.

**HOG and HOG-MHI Feature Descriptor for HFM Features:** In our system, like (Laptev *et al*., 2008), the local appearance 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. However, unlike

(Laptev *et al*., 2008), the motion features are represented by the HOG descriptors in the MHI (HOG-MHI). 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. In our approach, 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 $\mathrm{W}_{\min}$ (minimum window size) and $\mathrm{W}_{\max}$ (maximum window size). 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 $W_{min}$ and $W_{max}$. In our experiments, we set $\mathrm{W}_{\min} = 24$, $\mathrm{W}_{\max} = 48$. Our experiments demonstrate that using randomly selected window sizes handles scale variations very well and achieves better results than using fixed set of scales.

## ACTION DETECTION BY COMBINEING GAUSSIAN MIXTURE MODELS WITH BRANCH-AND-BOUND SEARCH

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

$$P_r(x^m|\theta^m) = \sum_{k=1}^{K} w^m(k)\, N(x^m;\ \mu^m(k), \textstyle\sum^m(k\ )) \qquad (3)$$

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

The advantages of GMM are that it is based on a well-understood statistical model, and it is easy to combine multiple features using GMMs. With GMM, we can estimate the probability that each feature vector $x^m$ belongs to the background or the action of interest. Suppose there are *C* categories of actions with parameter of $\Theta_1, \Theta_2, \dots, \Theta_C$. Each category corresponds to GMMs with *M* features $\Theta_C = \theta_C^1$, $\theta_C^2, \dots, \theta_C^M$.

The parameters of GMM can be estimated using maximum likelihood estimation. For example, for the *c* category, we first collect the subvolumes $V^c$ containing the action *c*, and then estimate GMMs parameters by maximizing the likelihood. A straightforward way is to independently train the model for each category and each feature. However, as shown by Reynolds (Reynolds *et al*., 2000), it is more effective to obtain $\theta_1^m,\ \theta_2^m, \dots, \theta_C^m$ coherently by the use of a universal background model. Following (Reynolds *et al*., 2000), we first train a background model $\theta_0^m$ which is independent to all the vectors $X^{all}$ using the *m* feature. Then we adapt $\theta_1^m, \dots, \theta_C^m$ from $\theta_0^m$ by *EM* algorithm in the following way.

We first estimate posterior probability of each $x_i^m$ subject to the background model $\theta_0^m$ by

$$p_k^c\left(x_p^m\right) = \frac{w(k)N(x_p^m, \mu_0^m(k), \sum_0^m(k\,))}{\sum_j w(j)N(x_p^m, \mu_0^m(j), \sum_0^m(j\,))} \qquad (4)$$

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

$$\mu_C^m(k) = \frac{1}{n_c}\sum_{x_p^m\in X^c} p_k^c\left(x_p^m\right) x_p^m \qquad (5)$$

Although we can update $\sum_c$ based on $p_k^c\left(x_p^m\right)$, in practice we force $w_C^m(k) = w_0^m(k)$ and $\sum_c^m(k\,) = \sum_0^m(k\,)$, which is computationally robust.

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

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

$$c^* = arg\max_c \sum_{m=1}^M \sum_{x_p^m\in V} \log P_r\left(x_p^m\middle|\theta_c^m\right) \qquad (6)$$

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

$$Q^* = arg\max_{Q\subseteq V} \mathcal{L}(Q|\Theta_c) = arg\max_{Q\subseteq V} \sum_m \sum_{p\in V} \log P_r\left(x_p^m\middle|\theta_c^m\right) \quad (7)$$

By assigning each patch a score $f\left(x_p^m\right) = \log P_r\left(x_p^m\middle|\theta_c^m\right)$, Equation (7) can be solved by branch-and-bound algorithm (Lampert *et al*., 2008, Yao and Zhu, 2009]. Branch-and-bound approach was first developed for integer programming problems. In recently years, it has been shown be to an efficient technique for object detection in images and action detection in videos (Lampert *et al*., 2008, Yao and Zhu, 2009, Yuan *et al*., 2009). Lampert *et al*. (Blaschko and Lampert, 2008, Lampert *et al*., 2008) showed that branch-and-bound can be used for object detection in 2D image base on a smart formulation. Yuan (Yuan *et al*., 2009) developed an efficient algorithm which generalizes branch-and-bound algorithm to 3D space of videos. In this chapter, we perform max-subvolume search using the 3D branch-and-bound algorithm in (Yuan *et al*., 2009), which is an extension of the 2D branch-and-bound technique (Lampert *et al*., 2008). The detailed technical description of the 3D branch-and-bound algorithm is omitted due to limited space.

## EXPERIMENTAL RESULTS OF ACTION DETECTION

## 1. Datasets

Our action detection scheme is evaluated by two datasets: KTH dataset (Schuldt *et al.*, 2004) and MSR Action Dataset II (MSRdataset, 2010).

**KTH Dataset**: The KTH dataset (Schuldt *et al.*, 2004) 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.

**MSR Action Dataset II**: MSR Action Dataset II is collected in Microsoft Research Redmond which was named MSR Action Dataset II, with cluttered background and multiple people move around. We do not use the CMU action dataset (Ke *et al*., 2007) since there is only a single sequence for training in it. Hu (Hu *et al.*, 2009) used videos from retailing surveillance, however, the dataset is confidential due to the privacy issue. Wang (Wang *et al*., 2009) collected a dataset of social game events, but their problem is about classification but not detection. MSR Action Dataset II includes 54 video sequences, each of which contains several different actions, e.g., hand waving, clapping, and boxing. These videos are taken with the background of parties, outdoor traffic, and walking people. Actors are asked to walk into the scene, perform one of the three kinds of action, and then walk out of the scenes with these backgrounds. Figure 1 shows the differences between KTH dataset (Figure 1(a)) and MSR Action Dataset II (Figure 1(b)). Note that in MSR Action Dataset II dataset there are many people in the scene and we need to locate the persons with actions of interest from the scene.

## 2. Action Detection Results on MSR Action Dataset II

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

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

$$(8)$$

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

where $| \cdot |$ denotes for the area of the subvolume, and $\delta_1, \delta_2$ are parameters to judge the overlapping ratio. In this chapter, both $\delta_1$ and $\delta_2$ are set as *1/4*.

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

$$Recall = \frac{\sum_{j=1}^{n} TD(Q_j^d)}{n}$$

$$Precision = \frac{\sum_{i=1}^{m} HG(Q_i^g)}{m} \qquad (9)$$

Where $n$ is the number of groundtruth and $m$ is the number of detected bounding boxes.

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

In MSR-II dataset, we use half of the videos for training and the remaining half videos for testing. We compare the detection results of each of the three features (Dollar *et al*., 2005, Laptev *et al*., 2008, Tian *et al*., 2011), and find that Hierarchically Filter Motion features outperform both Laptev's features (Laptev *et al*., 2008) and Dollar’s features (Dollar *et al*., 2005). We observe that both HFM and Laptev features can obtain reasonable detection results, while Dollar's features (Dollar *et al*., 2005) lead to results at a very low detection rate. The reason for the failure of Dollar's features might be that the Gabor filter based features are heavily affected by the cluttered background, since most of the detected patches fall in the background instead of action of interests. Since Dollar's features fail to detect some actions, we only compare results of two single feature detections and the multiple feature detection using our model. Figure 3 shows the precision-recall curves of action detection by using Hierarchically Filtered Motion features and Laptev STIP features respectively, and the fusions of multiple types of STIPs features. It can be seen that hierarchically filtered motion feature works better than Laptev's in handclapping and boxing, but comparable in handwaving. However, combining these two types of features, our multiple feature-based action detection schema works significantly better than using any single features in all the three actions. It is also interesting to see that if we incorporate the inappropriate features such as Dollar's features, the corresponding detection rate will decrease. The results confirm that *combining multiple relevant features will significantly improve the detection, while combining irrelevant feature might decrease the results*.

Figure 4 demonstrates some example results where the Hierarchically Filtered Motion features successfully detect the action of interests while Laptev’s STIP features fail. For each example, the top picture illustrates the detection results of using Hierarchically Filtered Motion features, while the bottom shows the results of using Laptev’s STIP features. The three colors denote different kinds of actions: red for clapping, green for waving, and blue for boxing. The proposed Hierarchically Filtered Motion features are more robust than other 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 Hierarchically 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.

Figure 5 shows the action detection results using our multiple feature model. Even the background is cluttered and there are multiple persons in both close and distant view, our detector works well and can locate the action of interest very accurately. Moreover, our detector is robust subject to short-term occlusions. Figure 6 shows the detection results with heavy occlusions.

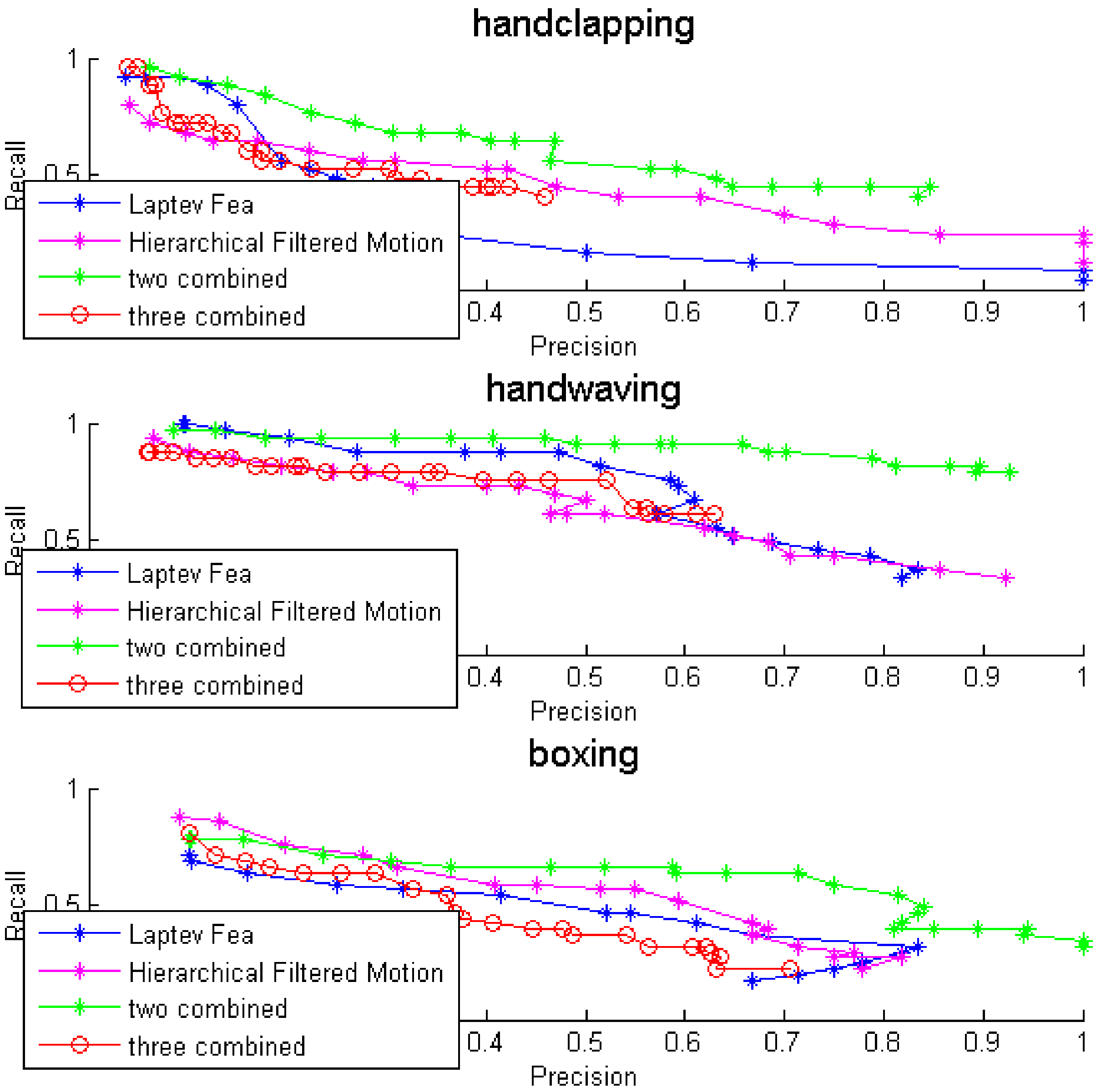


*Figure 3. Precision-Recall curves for MSR Action Dataset II for action detection by using HFM features and Laptev STIP features respectively, and the fusions of multiple types of STIPs features. The best results for all the 3 actions are achieved by combining HFM features and Laptev STIP features. The detection accuracy decreases by incorporating Dollar's features.*

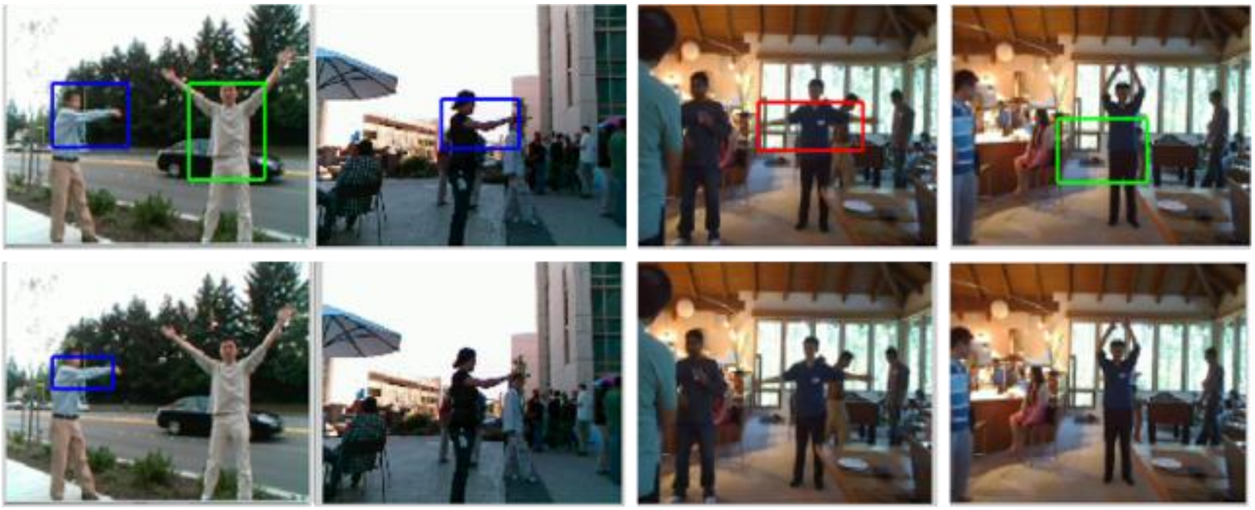


*Figure 4. Examples where HFM features successfully detect the action of interests while* Laptev's *STIP features fail. For each pair of examples, the picture in the top row illustrates the detection results of using HFM features, while the picture in the bottom row shows the result of using* Laptev's *STIP features. The three colors denote different kinds of actions: red for clapping, green for waving, and blue for boxing.*

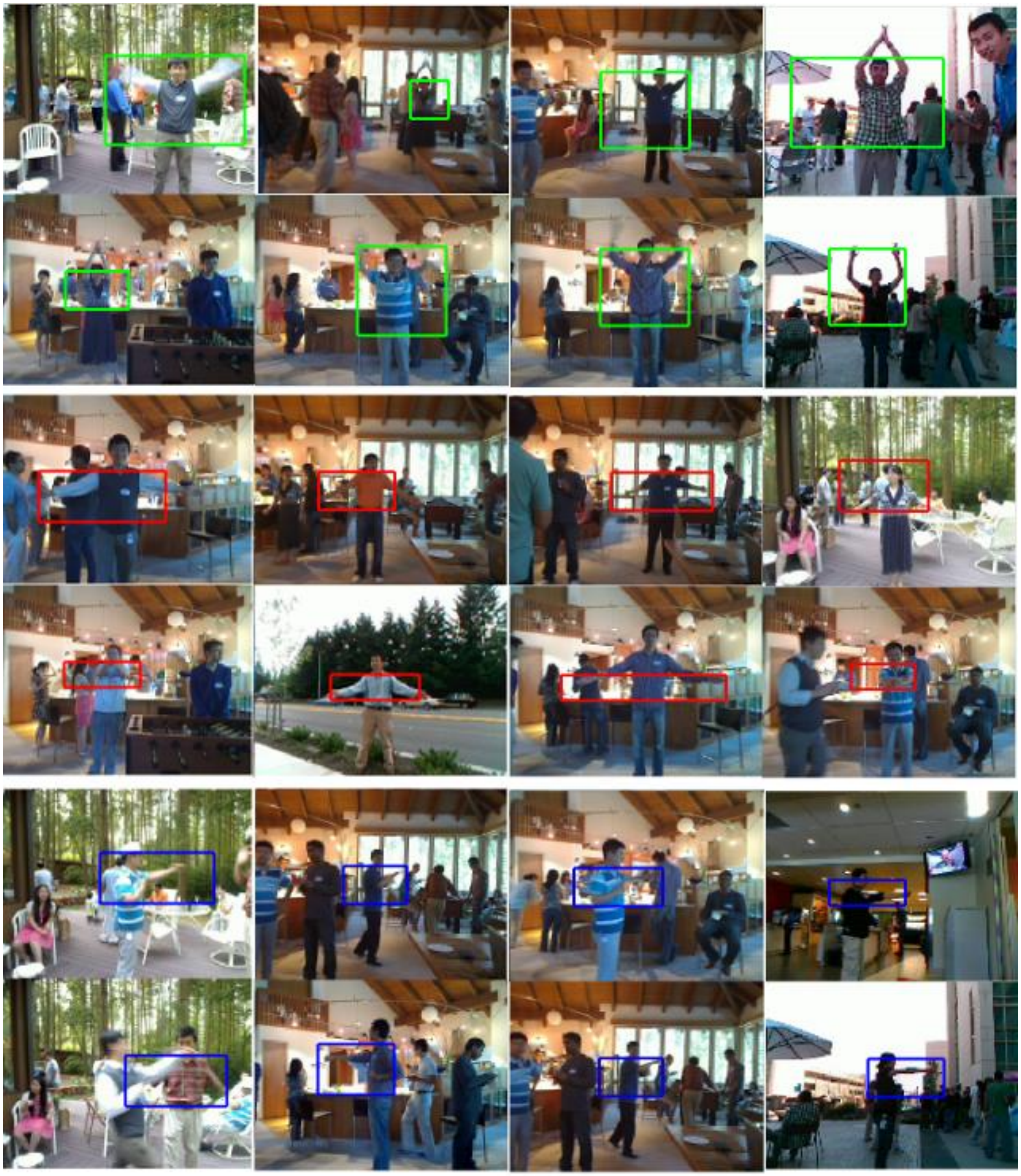


*Figure 5. Detection examples of MSR Action Dataset II. The bounding boxes denote the detected location using Branch-and-Bound search. The color of the bounding box denotes the action category: red for hand clapping, green for hand waving, and blue for boxing.*

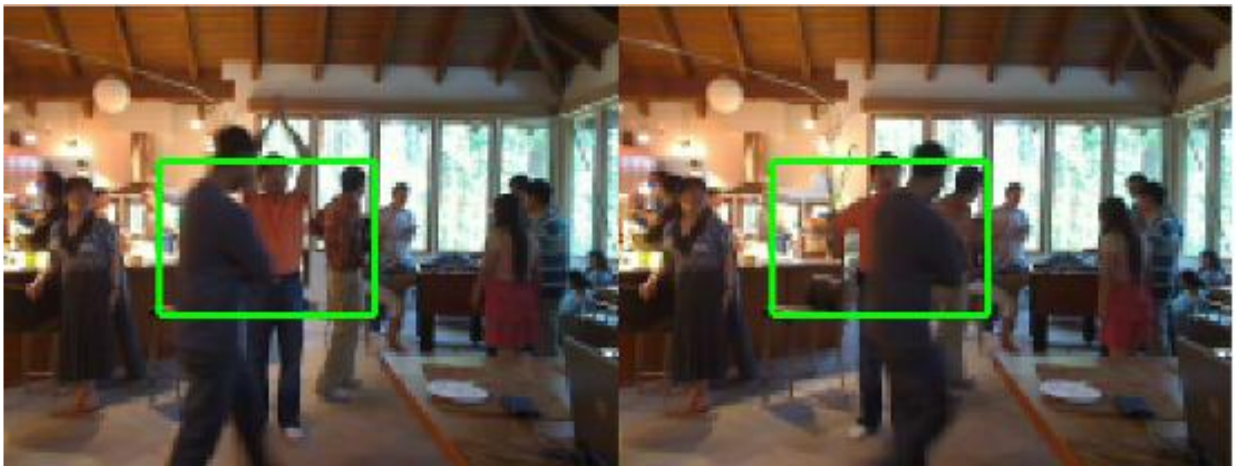


*Figure 6. Our detector successfully detects the action even with heavy occlusion.*

## 3. Action Detection Results on KTH Dataset

To compare our method with previous work, we test our algorithm on the public KTH dataset (Schuldt *et al*., 2004). In KTH dataset, each video sequence exhibits one individual action from beginning to end, locating the actions of interest is trivial. In each video of the KTH dataset, we need not estimate *Q* since there is only one actor repeating the same action without background motions involved, and all the STIPs in the video are associated with the action. However, the classification task on KTH dataset can still show how our multiple feature fusion method outperforms single feature based methods. To make the results comparable, we apply exactly the same experimental setting of KTH dataset as in (Schuldt *et al*., 2004). Among the 25 persons, we use 16 persons (1528 sequences) for training and the other 9 persons (863 sequences) for testing. Our method estimates the label of each video clip by Equation (8). Table 1 summarizes the action classification results from different types of features on KTH dataset. The results demonstrate that our feature fusion method outperforms the single feature classification results.

| **Method** | **Accuracy** |
|---|---|
| Schuldt *et al.*, 2004 | 71.71% |
| Dollar *et al*., 2005 | 80.7% |
| Yin *et al*., 2010 | 82% |
| Niebles *et al*., 2006 | 83.92% |
| Kaaniche *et al*., 2010 | 90.57% |
| JHuang *et al., 2007* | 91.6 |
| Laptev *et al.*, 2008 | 91.8% |
| Mikolajczyk e*t al., 2008* | 93.2% |
| Yuan *et al.*, 2009 | 93.3% |
| Kovashka *et al.*, 2010 | 94.53% |
| **Our method** | **94.5%** |

*Table 1.Comparison with the state-of-the art results on the KTH action dataset (6 actions with clean background)*

## 4. Computation Cost for HFM Feature Extraction

The proposed HFM feature extraction is very efficient. The computational costs for the following steps are evaluated on MSR Action Dataset II: 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. The average number of detected interest points is about 30 per image for each tested sequence. The details of the efficiency of the proposed Hierarchically Filtered Motion feature extraction are listed in Table 2. For the sequence with resolution of 160x120, the speed of interest point detection (step 1) is 216 frames per second. The speed for Hierarchically 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.

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

*Table 2. Computation cost for Hierarchically Filtered Motion feature extraction on MSR Action Dataset II with crowded background.*

## CONCLUSION

In this chapter, we have presented a new feature, Hierarchically Filtered Motion (HFM), for action detection in crowded videos without tracking objects or key points. The HFM features can reduce distracting motions caused by the background moving objects near an interest point. The proposed HFM is more robust than other STIP features for action recognition in crowded videos. We further build a novel framework which combines GMM-based representation of HFM with other STIPs and branch-and-bound based detection. We have performed action detection experiments on MSR Action Dataset II for videos with cluttered and moving background and KTH dataset. Experiment results have demonstrated that our approach outperforms existing techniques and can effectively detect actions even with cluttered background and partial occlusions.

## REFERENCES

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

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

Bobick, A. and Davis, J. (2001), The recognition of human movement using temporal templates. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 23, 257–267.

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

Cao, L., Liu, Z., and Huang, T. (2010), *Cross-dataset action detection*. IEEE Conference on Computer Vision and Pattern Recognition.

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

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

Dalal, N., and Triggs, B. (2005), *Histograms of Oriented Gradients for Human Detection*, IEEE International Conference on Computer Vision.

Davis, J. (2001) "*Hierarchical motion history images for recognizing human motion*", Proc. Of IEEE Workshop on Detection and Recognition of Events in Video.

Dollar, P., Rabaud, V., Cottrell, G., and Belongie, S. (2005), *Behavior recognition via sparse spatio-temporal features*. IEEE International Workshop on VS-PETS.

Du, T., and Yuan, J. (2011), *Optimal Spatio-Temporal Path Discovery for Video Event Detection*, IEEE Conference on Computer Vision and Pattern Recognition.

Harris, C. and Stephens, M. (1988), *A combined corner and edge detector*. In Proceeding of Alvey Vision Conference, pages 189–192.

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

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

Ji, X., and Liu, H. (2010), 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.*

Junejo, I., Dexter, E., Laptev, I., and Perez,P. (2011), 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.*

Kaaniche, M. and Bremond, F. (2010), Gesture Recognition by Learning Local Motion Signatures, IEEE International Conference on Computer Vision.

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

Kovashka, A., and Grauman, K. (2010), *Learning a Hierarchy of Discriminative Space-Time Neighborhood Features for Human Action Recognition*, IEEE International Conference on Computer Vision.

Lampert, C. Blaschko, M.and Hofmann, T. (2008), *Beyond sliding windows: Object localization by efficient subwindow search*. IEEE Conference on Computer Vision and Pattern Recognition.

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

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

Liu, J., S. Ali, S., and Shah, M. (2008), Recognizing human actions using multiple features. IEEE Conf. on Computer Vision and Pattern Recognition.

Liu, J., Luo, J., and Shah, M. (2009), *Recognizing realistic actions from videos "in the wild"*. IEEE Conference on Computer Vision and Pattern Recognition.

M. Marszałek, M., Laptev, I., and Schmid. C. (2009), *Actions in context*. IEEE Conference on Computer Vision and Pattern Recognition.

Mattivi R. and L. Shao, L. (2009), Human Action Recognition Using LBP-TOP as Sparse Spatio-Temporal Feature Descriptor, *Computer Analysis of Images and Patterns*.

Messing, R., Pal, C., and Kauze, H. (2009), Activity Recognition Using the Velocity Histories of Tracked Keypoints, IEEE Conference on Computer Vision.

Mikolajczyk K., and Uemura, H. (2008), *Action recognition with motion-appearance vocabulary forest*. IEEE Conference on Computer Vision and Pattern Recognition.

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

Pang, Y., Yuan, Y., Li, X., and Pan, J. (2010), Efficient HOG human detection, *Signal Processing* (15 September 2010) http://dx.doi.org/ 10.1016/j.sigpro.2010.08.010.

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

Rodriguez, M., Ahmed, J. and Shah, M. (2008), *Actionmach: A spatio-temporal maximum average correlation height filter for action recognition*. IEEE Conference on Computer Vision and Pattern Recognition.

Schuldt, C., Laptev, I. and Caputo, B. (2004), *Recognizing human actions: A local SVM approach*. International Conference on Pattern Recognition.

Sun, J., Wu, X., Yan, S., Cheong, L., Chua, T., and Li, J. (2009), *Hierarchical Spatio-Temporal Context Modeling for Action Recognition*, IEEE Conf. on Computer Vision and Pattern Recognition.

Tian, Y., Cao, L., Liu, Z., and Zhang, Z. (2011), Hierarchical filtered motion field for action recognition in crowded videos. *IEEE Transactions on Systems, Man, and Cybernetics--Part C: Applications and Reviews*.

Wang, P., Abowd, G. and Rehg, J. (2009), *Quasi-periodic event analysis for social game retrieval*. In IEEE International Conference on Computer Vision.

Wong, S., Kim, T., and Cipolla, R. (2007), *Learning motion categories using both semantic and structural information*. IEEE Conference on Computer Vision and Pattern Recognition.

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

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

Yin, J., and Meng, Y. (2010), *Human Activity Recognition in Video using a Hierarchical Probabilistic Latent Model*, IEEE Conf. on Computer Vision and Pattern Recognition.

Yuan, Y., and Pang, Y. (2008), Discriminant *Adaptive Edge Weights for Graph Embedding*," IEEE International Conference on Acoustics, Speech, and Signal Processing.

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

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

MSR Action Dataset II, (2009), http://research.microsoft.com/en-us/um/people/zliu/ActionRecoRsrc/.

## KEY TERMS & DEFINITIONS

**Human Action Recognition**: *classify a video to a predefined action category*

**Human Action Detection**: *not only answer "which action happens in a video", but also "when and where the action happens in the video sequence.*

**Spatiotemporal Interest Point**: *features of local structures in space-time where the image values have significant local variations in both space and time*.

**Hierarchically Filtered Motion**: *spatiotemporal motion features with global and local filters*.

**Motion History Image:** *a real-time motion template that temporally layers consecutive image differences into a static image template.*

**Histograms of Oriented Gradient**: *feature descriptors that count occurrences of gradient orientation in localized portions of an image*.

**Branch-and-Bound Algorithm**: *a general algorithm for finding optimal solutions of various optimization problems, especially in discrete and combinatorial optimization*.