# A Human Action Recognition System for Embedded Computer Vision Application

Hongying Meng, Nick Pears, Chris Bailey
Department of Computer Science
The University of York, Heslington, York,YO10 5DD,UK
{hongying,nep,chrisb}@cs.york.ac.uk

## Abstract

*In this paper, we propose a human action recognition system suitable for embedded computer vision applications in security systems, human-computer interaction and intelligent environments. Our system is suitable for embedded computer vision application based on three reasons. Firstly, the system was based on a linear Support Vector Machine (SVM) classifier where classification progress can be implemented easily and quickly in embedded hardware. Secondly, we use compacted motion features easily obtained from videos. We address the limitations of the well known Motion History Image (MHI) and propose a new Hierarchical Motion History Histogram (HMHH) feature to represent the motion information. HMHH not only provides rich motion information, but also remains computationally inexpensive. Finally, we combine MHI and HMHH together and extract a low dimension feature vector to be used in the SVM classifiers. Experimental results show that our system achieves significant improvement on the recognition performance.*

## 1. Introduction

Event detection in video is becoming an increasingly important computer vision application, particularly in the context of activity classification [1]. Event recognition is a fundamental building block for interactive systems which can respond to gestural commands and for systems which analyse body motion, for example in sporting activities and dance. Other applications include video surveillance and video indexing.

Aggarwal and Cai [1] present an excellent overview of human motion analysis. Of the appearance based methods, template matching has gained increasing interest recently [18, 19, 10, 20, 7, 16, 5, 2, 17, 12, 11, 21, 22, 15]. These methods are based on the extraction of a 2D or 3D shape model directly from the images, to be classified (or matched) against a training data. Motion-based models do not rely on static models of the person, but on human motion characteristics. Motion feature extraction is the key component in these kinds of human action recognition systems.

In this paper, we aim to build a human action recognition system based on a linear Support Vector Machine (SVM) classifier and compact motion features from the videos with a view to applications in security systems, human-computer interaction and intelligent environments.

The rest of this paper is organized as follows: In section 2, we give an introduction to related work. In section 3, we give a detailed description of our new Hierarchical Motion History Histogram (HMHH) features. In section 4, we give a brief overview of our SVM-based human action recognition system. In section 5, experimental results are presented and compared and finally, we present conclusions.

## 2. Related work

Bobick and Davis [3] pioneered the idea of temporal templates [14]. They use Motion Energy Images (MEI) and MHI to recognize many types of aerobics exercises. They [4] also proposed the Motion Gradient Orientation (MGO) to explicitly encode changes in an image introduced by motion events.

Davis [6] also presented a useful hierarchical extension for computing a local motion field from the original MHI representation. The MHI was transformed into an image pyramid, permitting efficient fixed-size gradient masks to be convolved at all levels of the pyramid, thus extracting motion information at a wide range of speeds. The hierarchical MHI approach remains a computationally inexpensive algorithm to represent, characterize, and recognize human motion in video.

Schuldt et al. [18] proposed a method for recognizing complex motion patterns based on local space-time features in video and they integrated such representations with SVM classification schemes for recognition.

1-4244-1180-7/07/$25.00 ©2007 IEEE

The work of Efros et al. [8] focuses on the case of low resolution video of human behaviours, targeting what they refer to as the 30 pixel man. In this setting, they propose a spatio-temporal descriptor based on optical flow measurements, and apply it to recognize actions in ballet, tennis and football datasets.

Weinland et al. [19] introduced Motion History Volumes (MHV) as a free-viewpoint representation for human actions in the case of multiple calibrated, and background-subtracted, video. They presented algorithms for computing, aligning and comparing MHVs of different actions performed by different people from a variety of viewpoints.

Ke et al. [10] studied the use of volumetric features as an alternative to the local descriptor approaches for event detection in video sequences. They generalized the notion of 2D box features to 3D spatio-temporal volumetric features. They constructed a real-time event detector for each action of interest by learning a cascade of filters based on volumetric features that efficiently scanned video sequences in space and time.

Ogata et al. [16] proposed Modified Motion History Images (MMHI) and used an eigenspace technique to realize high-speed recognition. The experiment was performed on recognizing six human motions.

Wong and Cipolla [20] proposed a new method to recognize primitive movements based on MGO extraction and used it for continuous gesture recognition [21] later.

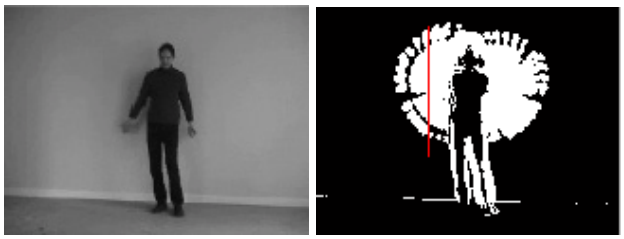
Recently, Dalal et al. [5] proposed Histogram of Oriented Gradient (HOG) appearance descriptors for image sequences and developed a detector for standing and moving people in video.

Dollár et al. [7] proposed a similar method where they use a new spatio-temporal interest point detector to obtain a global measurement instead of the local features in [8]. Niebles et al. [15] also use spatial-time interest points to extract spatial-temporal words as their features. Yeo et al. [22] estimate motion vectors from optical flow and calculate frame-to-frame motion similarity to analyse human action in video.

Blank et al. [2] regarded human actions as three dimensional shapes induced by silhouettes in space-time volume. They adopted an approach for analyzing 2D shapes and generalized it to deal with volumetric space-time action shapes.

Oikonomopoulos et al. [17] introduced a sparse representation of image sequences as a collection of spatio-temporal events that were localized at points that were salient both in space and time for human action recognition.

We note that, in some of these methods, the motion features employed are relatively complex [8, 18, 19, 15, 5, 7, 2, 17, 10, 22], which implies significant computational cost when building the features. Some of them require segmentation, tracking or other prohibitive computational cost processes [3, 4, 6, 20, 21, 16, 2], which makes them not suitable for real-time embedded vision applications in the intelligent environment.



(a) Handwaving sample (b) MHI of Handwaving

Figure 1. Example of a MHI. (a) is one frame from the original hand waving action video clip and (b) is the MHI of this action. The vertical red line in (b) has the pixels from $(60, 11)$ to $(60, 80)$.

In our previous work[12, 11, 13], we have proposed a SVM based system based on the simple motion features MHI,MMHI and MGO. They are suitable for embedded computer vision application, but the overall performance on real-world (challenging) databases is relatively poor.

In this work, we aim for a solution which uses compact representations, is fast to compute, and yet gives an improved classification performance over existing compact and fast methods. We extended the work of [12, 11, 13] by introducing new motion features and combination methods with significiently improved performance.

## 3. Hierarchical Motion History Histogram (HMHH)

In this section, we introduce the "Hierarchical Motion History Histogram" (HMHH) and describe the motivation for it, but first it is necessary to review the motion history image (MHI).

### 3.1. Motion History Image

A MHI [3] is a kind of temporal template to compact the whole motion sequence into one image to represent the motion. It is the weighted sum of past images and the weights decay back through time. Therefore, a MHI image contains the past raw images within itself, where most recent image is brighter than past ones.

Normally, a MHI $H_\tau(u, v, k)$ at time $k$ and location $(u, v)$ is defined by the following equation 1:

$$H_\tau(u,v,k) = \begin{cases} \tau, & if\ D(u,v,k) = 1 \\ max\{0, H_\tau(u,v,k) - 1\}, & otherwise \end{cases} \quad (1)$$

where the motion mask $D(u, v, k)$ is a binary image obtained from subtraction of frames, and $\tau$ is the maximum duration a motion is stored. In general, $\tau$ is chosen as the constant 255, allowing the MHI to be easily represented as a

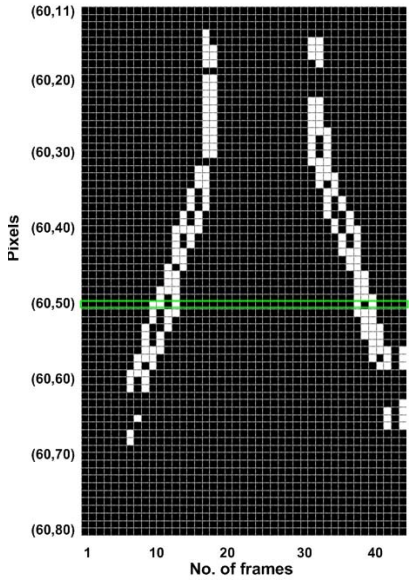



Figure 2. $D(:,:,:)$ on the red line of figure 1(b) is shown. Each row is $D(u,v,:)$ for one fixed pixel $(u,v)$. A white block represents '1' and a black block '0'. The horizontal green line is the 'binarised frame difference history' or 'motion mask' of pixel $(60,50)$ through time, ie, $D(60,50,:)$.

gray scale image with one byte depth. Thus a MHI pixel can have a range of values, whereas a MEI is its binary version, which can easily be computed by thresholding $H_\tau > 0$.

An example of a MHI is shown in Figure 1, where (a) is one frame from the original hand waving action video clip and (b) is the MHI of this action. It is clear that MHI is an image, with the same size as the frame, but which retains some motion information of the action.

In order to have a detailed look at the MHI, we have selected the pixels on the red line in the MHI of figure 1 (b). If some action happened at frame $k$ on pixel $(u,v)$, then $D(u,v,k)=1$, otherwise $D(u,v,k)=0$. The locations of these pixels are $(60,11),(60,12),\cdots,(60,80)$. For a pixel $(u,v)$, the motion mask $D(u,v,:)$ of this pixel is the binary sequence:

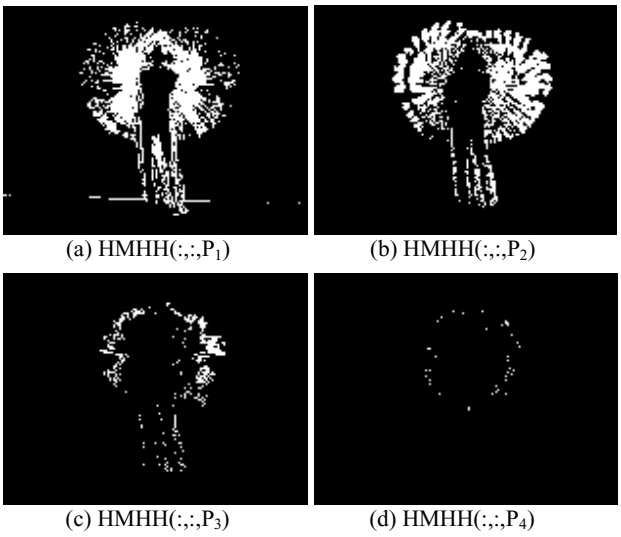
$$D(u,v,:)=(b_1,b_2,\cdots,b_N)\,,\;\; b_i\in\{0,1\} \qquad (2)$$

where $N+1$ is the total number of frames.

All of the motion masks on the red line are shown in figure 2. Each row is $D(u,v,:)$ for one fixed pixel $(u,v)$ and a white block represents '1' and black block represents '0' in the sequences. The green line is the motion mark $D(60,50,:)$ and it has the following sequence 3:

$$00000000011010000000000000000000000001010000 \quad (3)$$

From the definition of MHI in equation 1 it can be observed that, for each pixel $(u,v)$, MHI only retains the most recent action that occurred. That is, only the last '1' in the sequence 3 are retained in the MHI at pixel $(60,50)$. It is clear that previous '1' in the sequence, when some action occurred, are not represented. It is also clear that almost all the pixels have more than one '1' in their sequence. This has motivated us to design a new representation (the HMHH, described in the next sub-section) in which all of the information in the sequence is used and yet it remains compact and efficient to use.



(a) HMHH(:,:,$P_1$) (b) HMHH(:,:,$P_2$)

(c) HMHH(:,:,$P_3$) (d) HMHH(:,:,$P_4$)

Figure 3. HMHH example. Four Patterns $P_1, P_2, P_3$ and $P_4$ were selected. The results were generated from the handwaving action in figure 1. Each pattern $P_i$, HMHH$(:,:,P_i)$ has the same size as the original frame.

### 3.2. HMHH

We define the patterns $P_i$ in $D(u,v,:)$ sequences based on the number of connected '1's as showed in equation 4.

$$\begin{aligned} P_1 &= 010 \\ P_2 &= 0110 \\ P_3 &= 01110 \\ &\vdots \\ P_M &= 0\underbrace{1\cdots1}_{M}0 \end{aligned} \qquad (4)$$

We denote a subsequence $C_i$ by equation 5 and denote the set of all subsequences of $D(u,v,:)$ as $\Omega\{D(u,v,:)\}$. Then for each pixel $(u,v)$, we can count the number of occurrences of each specific pattern $P_i$ in the sequence $D(u,v,:)$, as shown in equation 6, where $\mathbf{1}$ is the indicator function.

$$C_i = b_{n_1}, b_{n_2}, \cdots, b_{n_i} \qquad (5)$$

$$HMHH(u,v,P_i) = \textstyle\sum_j \mathbf{1}_{\{\mathbf{C_j}=\mathbf{P_i}|\mathbf{C_j}\in\mathbf{\Omega}\{\mathbf{D(u,v,:)}\}\}} \qquad (6)$$

From each pattern $P_i$, we can build a gray scale image and we call this the Motion History Histogram (MHH), since the bin value records the number of this pattern type. With all the patterns $P_i, i = 1...M$ together, we collectively call them the 'Hierarchical Motion History Histogram' (HMHH) representation.

For a pattern $P_i$, HMHH$(:,:,P_i)$ can be displayed as an image. In figure 3, four patterns $P_1, P_2, P_3$ and $P_4$ are shown, which were generated from the handwaving action in figure 1. By comparing the HMHH in figure 3 with the MHI in figure 1, it is interesting to find that HMHH decomposes the region of MHI into different parts based on patterns. Unlike the hierarchical MHI described by Davis [6], where only small size MHIs were obtained, HMHH records the rich information of an action. The computation of HMHH is inexpensive and can be implemented in the following procedure.

---

**Algorithm (HMHH)**

---

**Input:** Video clip *f (u,v,k), u=1,...,U, v=1,...,V,* frame *k=0,1,...,N*
**Initialisation**: Pattern *M,* HMHH*(1:U,1:V,1:M)=0*, *I(1:U,1:V)=1*
**For** *k=1* to *N* (For 1)
  **Compute**: *D(:,:,k)*
  **For** *u=1 to U* (For 2)
    **For** *v= 1 to V* (For 3)
      **If** Subsequence $C_j$=*{D(u,v,I),...,D(u,v,k)}*=$P_i$
        **Update:** HMHH*(u,v,*$P_i$*)*=HMHH*(u,v,*$P_i$*)+1*
      **End** *If*
      **Update:** *I(u,v)*
    **End** (For 3)
  **End** (For 2)
**End** (For 1)
**Output**: HMHH*(1:U,1:V,1:M)*

---

### 3.3. Motion Geometric Distribution (MGD)

The size of the HMHH representation is a bit large to present to a classifier and we seek a more compact representation, which captures the geometric distribution of the motion across the image. To do this, we first define the binary version of a MHH as $MHH_b$, as shown equation 7.

$$MHH_b(u,v,P_i) = \begin{cases} 1, & if\ MHH(u,v,P_i) > 0 \\ 0, & otherwise \end{cases} \quad (7)$$

We then sum each row of $MHH_b$ (for a given pattern, $P_i$) to give a vector of size $V$ rows. We obtain another vector by summing columns to give another vector of size $U$ rows. Thus using all $M$ levels in the binarised MHH hierarchy, we obtain a 'Motion Geometric Distribution' (MGD) vector of size $M \times (U + V)$, which is relatively compact, when compared to the size of the original HMHH and MHI features. The MGD vector can thus be represented by the following equation 8:

$$MGD = \{\sum_u MHH_b(u,v,P_i), \sum_v MHH_b(u,v,P_i)\} \quad i = 1, 2, \cdots, M \quad (8)$$

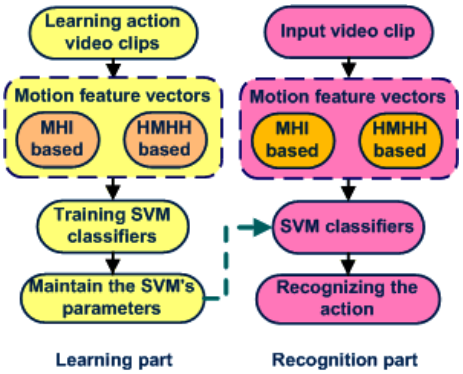


Figure 4. SVM based human action recognition system.

## 4. SVM based human action recognition

Meng et al. [12] built a fast human action recognition system based on a linear SVM and simple motion features. In this architecture, a linear SVM was chosen because it has historically shown very good performance in lots of real-world classification problems and also can deal with very high dimensional feature vectors. Three fundamental motion features MHI, MMHI and MGD were tested in the system with varying levels of performance.

We make a generalization for the system in [12] , allowing it to handle more motion features. The overall architecture of the human action system is shown in figure 4. The training part is done off-line to get SVM parameters. The classification part is just a inner product between SVM parameters and motion feature obtained from testing video.

## 5. Experimental results

For the evaluation, we use a challenging human action recognition database, recorded by Christian Schuldt [18]. It contains six types of human actions (walking, jogging, running, boxing, hand clapping and hand waving) performed several times by 25 subjects in four different scenarios: outdoors (s1), outdoors with scale variation (s2), outdoors with different clothes (s3) and indoors (s4). This database contains 2391 sequences. Figure 5 shows the examples in each type of human action and their associated MHI and HMHH motion features.

We performed our experiments in the same manner as in papers [10, 12, 11, 13]. In all our experiments, the same

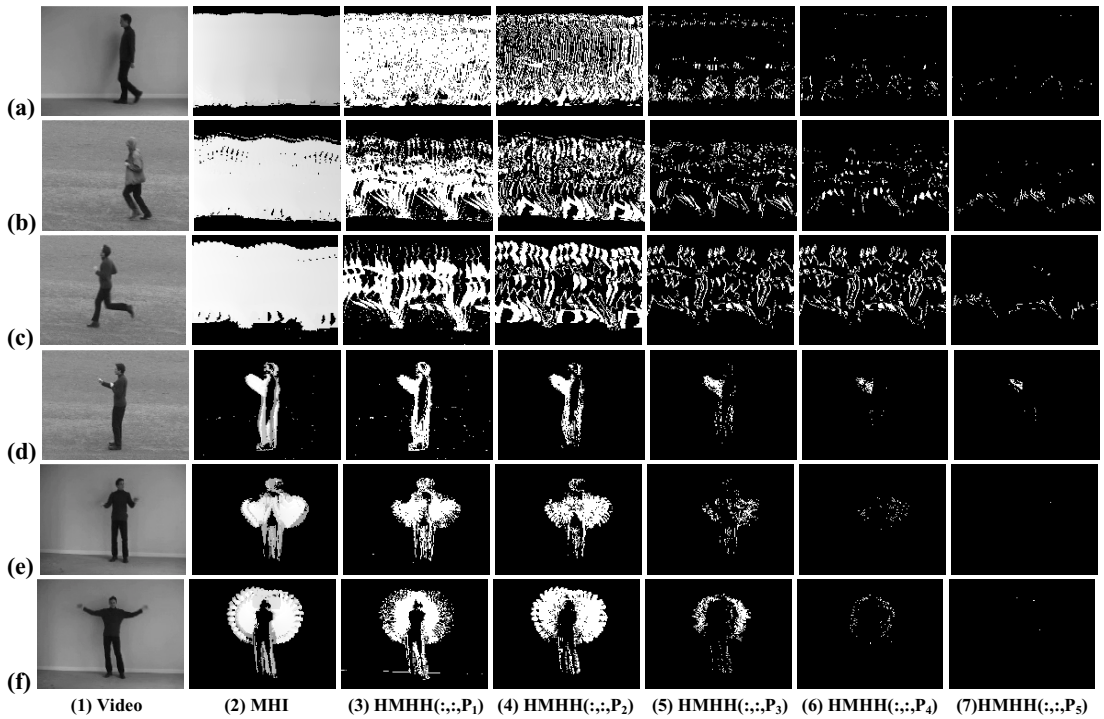



Figure 5. The six database human actions and associated MHI, HMHH features: (a) walking (b) jogging (c) running (d) boxing (e) handclapping (f) hand-waving.

parameters were used. The threshold in frame differencing was chosen as 25 and $\tau$ was chosen as 255 for MHI construction. The number of the patterns was chosen to be $M = 5$ for HMHH computation. The size of the MHI is $160 \times 120 = 19200$, which is same width as that of the frames in the videos. In our experiment, the SVM is implemented using the $SVM^{light}$ software [9].

We experimented with different HMHH and MHI based features in our system. In order to avoid a very high dimensional HMHH, we constructed a small sized HMHH_S by averaging the pixels in an $8 \times 8$ block, so that the size of all HMHH feature vectors is reduced to $20 \times 15 \times 5 = 1500$. Our MGD feature also has a small size of $(160+120) \times 5 = 1400$. In [13], a histogram of MHI was combined with Haar Wavelet Transform (HWT) of MHI features and good results obtained, when compared to other approaches on the same dataset. Here, we combine the histogram of MHI and MDG, which has a size of $255 + 1400 = 1655$. Table 1 showed the confusion matrix. The confusion matrices show the motion label (vertical) versus the classification results (horizontal). Each cell $(i, j)$ in the table shows the percentage of class $i$ action being recognized as class $j$. Then trace of the matrices show the percentage of the correctly recognized action, while the remaining cells show the percentage of misclassification.

We compared our results with other methods on this challenging dataset and summarize the correctly classified rates in table 2. From this table, we can see that HMHH has made a significant improvement in comparison with MHI. Furthermore, MGD gives better performance than HMHH itself. The best performance, which gives significantly better classification results, came from the combined feature based on the histogram of MHI [13] and MGD.

| | Walk | Jog | Run | Box | Clap | Wave |
|---|---|---|---|---|---|---|
| Walk | **66.0** | 31.3 | 0.0 | 0.0 | 2.1 | 0.7 |
| Jog | 13.9 | **62.5** | 21.5 | 1.4 | 0.0 | 0.7 |
| Run | 2.1 | 16.7 | **79.9** | 0.0 | 0.0 | 1.4 |
| Box | 0.0 | 0.0 | 0.0 | **88.8** | 2.8 | 8.4 |
| Clap | 0.0 | 0.0 | 0.0 | 3.5 | **93.1** | 3.5 |
| Wave | 0.0 | 0.0 | 0.0 | 1.4 | 6.9 | **91.7** |

Table 1. MGD & Hist. of MHI's confusion matrix, trace=481.9

It should be mentioned here that some results avoid the difficult part of the dataset (subset 2, outdoor with scale variation) [7, 22] and some of them [18, 15, 22] did an easier task of classifying each complete sequence (containing 4 repetitions of same action) into one of six classes while our method was trained as the same way as papers [10, 8, 12, 11, 13]; that is to detect a single instance of each action within arbitrary sequences in the dataset.

| Method | Rate(%) |
|---|---|
| SVM on local features [18]$*$ | 71.7 |
| Cascade of filters on volumetric features[10] | 62.9 |
| SVM on MHI [12] | 63.5 |
| SVM_2K on MHI & MMHI [11] | 65.3 |
| SVM on HMHH_S | 69.6 |
| SVM on MGD | 72.1 |
| SVM on HWT of MHI & Hist. of MHI[13] | 70.9 |
| **SVM on MGD & Hist. of MHI** | **80.3** |
| SVM on spatio-temporal feature [7]$\Delta$ | 81.2 |
| Learning on spatial-temporal words [15] $*$ | 81.5 |
| KNN on NZMS [22]$\Delta*$ | 86.0 |

Table 2. Overall correctly classified rate (%) for all the methods on this open, challenging dataset. Some of them didn't use the difficut part of dataset($\Delta$) while some of them did an easier task($*$).

## 6. Conclusion

In this paper, we have addressed the limitations of the MHI representation and proposed a new HMHH feature, which keeps the whole historical motion information in the video. HMHH remains a computationally inexpensive feature to represent, characterize motion in video.

We extract a basic MGD feature vector from HMHH and apply it in an SVM based human action recognition system. In comparison with methods in [18, 10, 7, 15, 22], our feature vectors are computationally inexpensive. In comparison with methods in [12, 11, 13] methods, we used a similarly inexpensive and fast method, but we obtained a significant improvement on recognition performance.

For future work,we have developed a FPGA based real-time video system and the algorithms will be modified and optimized based on the hardware limitations such as memory, speed and storage space. We will also consider its potential applications in hand gesture recognition, facial expression classification and video characterization.

## References


[1] J. K. Aggarwal and Q. Cai. Human motion analysis: a review. *Comput. Vis. Image Underst.*, 73(3):428–440, 1999.

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, pages 1395–1402, 2005.

[3] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(3):257–267, 2001.

[4] G. R. Bradski and J. W. Davis. Motion segmentation and pose recognition with motion history gradients. *Mach. Vis. Appl.*, 13(3):174–184, 2002.

[5] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV (2)*, pages 428–441, 2006.

[6] J. W. Davis. Hierarchical motion history images for recognizing human motion. In *IEEE Workshop on Detection and Recognition of Events in Video*, pages 39–46, 2001.

[7] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, October 2005.

[8] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *ICCV*, pages 726–733, 2003.

[9] T. Joachims. Making large-scale SVM learning practical. In *Advances in Kernel Methods - Support Vector Learning*, USA, 1999. MIT-Press. Oikonomopoulos, Antonios and Patras, Ioannis and Pantic, Maja eds.

[10] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *ICCV*, pages 166–173, 2005. Beijing, China, Oct. 15-21, 2005.

[11] H. Meng, N. Pears, and C. Bailey. Human action classification using SVM_2K classifier on motion features. In *LNCS*, volume 4105, pages 458–465, Istanbul, Turkey, 2006.

[12] H. Meng, N. Pears, and C. Bailey. Recognizing human actions based on motion information and SVM. In *Proceedings of 2nd IET International Conference on Intelligent Environments*, pages 239–245, Athens, Greece, 2006. IET.

[13] H. Meng, N. Pears, and C. Bailey. Motion information combination for fast human action recognition. In *Proceedings of 2nd International Conference on Computer Vision Theory and Applications (VISAPP07)*, Barcelona, Spain., March 2007.

[14] T. Moeslund, A. Hilton, and V. Kruger. A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Underst.*, 103(2-3):90–126, November 2006.

[15] J. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. In *BMVC06*, page III:1249, 2006.

[16] T. Ogata, J. K. Tan, and S. Ishikawa. High-speed human motion recognition based on a motion history image and an eigenspace. *IEICE Transactions on Information and Systems*, E89(1):281–289, 2006.

[17] A. Oikonomopoulos, I. Patras, and M. Pantic. Kernel-based recognition of human actions using spatiotemporal salient points. In *Proceedings of CVPR workshop 06*, volume 3, pages 151–156, June 2006.

[18] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *ICPR*, Cambridge, U.K, 2004.

[19] D. Weinland, R. Ronfard, and E. Boyer. Motion history volumes for free viewpoint action recognition. In *IEEE International Workshop on modeling People and Human Interaction (PHI'05)*, 2005.

[20] S.-F. Wong and R. Cipolla. Real-time adaptive hand motion recognition using a sparse bayesian classifier. In *ICCV-HCI*, pages 170–179, 2005.

[21] S.-F. Wong and R. Cipolla. Continuous gesture recognition using a sparse bayesian classifier. In *ICPR (1)*, pages 1084–1087, 2006.

[22] C. Yeo, P. Ahammad, K. Ramchandran, and S. Sastry. Compressed domain real-time action recognition. In *IEEE International Workshop on Multimedia Signal Processing (MMSP) - 2006*, Washington, DC, USA, 2006.