Motion Feature Combination for Human Action Recognition in Video

Hongying Meng[1], Nick Pears[2], and Chris Bailey[2]

[1] Department of Computing and Informatics
University of Lincoln, Brayford Pool, Lincoln, LN6 7TS, U.K.
hmeng@lincoln.ac.uk
[2] Department of Computer Science, University of York, York, YO10 5DD, U.K.

Abstract. We study the human action recognition problem based on motion features directly extracted from video. In order to implement a fast human action recognition system, we select simple features that can be obtained from non-intensive computation. We propose to use the motion history image (MHI) as our fundamental representation of the motion. This is then further processed to give a histogram of the MHI and the Haar wavelet transform of the MHI. The combination of these two features is computed cheaply and has a lower dimension than the original MHI. The combined feature vector is tested in a Support Vector Machine (SVM) based human action recognition system and a significant performance improvement has been achieved. The system is efficient to be used in real-time human action classification systems.

Keywords: Event recognition, Human action recognition, Video analysis, Support Vector Machine.

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 an important goal for building intelligent systems which can react to what is going on in a scene. Event recognition is also a fundamental building block for interactive systems which can respond to gestural commands, instruct and correct a user learning athletics, gymnastics or dance movements, or interact with live actors in an augmented dance or theatre performance.

Recognizing actions of human actors from digital video is a challenging topic in computer vision with many fundamental applications in video surveillance, video indexing and social sciences. Feature extraction is the basis to perform many different tasks with video such as video object detection, object tracking and object classification.

Model based method are extremely challenging as there is large degree of variability in human behaviour. The highly articulated nature of the body leads to high dimensional models and the problem is further complicated by the non-rigid behaviour of clothing. Computationally intensive methods are needed for nonlinear modeling and optimisation. Recent research into anthropology has revealed that body dynamics are far more

J. Braz et al. (Eds.): VISIGRAPP 2007, CCIS 21, pp. 151–163, 2008.

complicated than was earlier thought, affected by age, ethnicity, gender and many other circumstances [2].

Appearance-based models are based on the extraction of a 2D shape model directly from the images, to be classified (or matched) against a trained one. Motion-based models do not rely on static models of the person, but on human motion characteristics. Motion feature extraction and selection are two of the key components in these kinds of human action recognition systems.

In this paper, we study the human action classification problem based on motion features directly extracted from video. In order to implement fast human action recognition, we select simple features that can be obtained from non-intensive computation. In particular, we use the Motion History Image (MHI) [3] as our fundamental feature. We propose novel extraction methods to extract both spatial and temporal information from these initial MHI representations and we combine them as a new feature vector that has a lower dimension and provides better motion action information than the raw MHI information. This feature vector was used in a Support Vector Machine (SVM) based human action recognition system.

The rest of this paper is organised as follows: In section 2, we will give an introduction to some related work. In section 3, we give a brief overview of our system. In section 4, the detailed techniques of this system are explained including motion features, feature extraction methods and SVM classifier. In section 5, some experimental results are presented and compared. In section 6, the same combination idea has been tested on other features and significant improvement is also achieved. Finally, we give the conclusions.

2 Previous Work

Aggarwal and Cai [1] present an excellent overview of human motion analysis. Of the appearance based methods, template matching has increasingly gained attention. Bobick and Davis [3] use Motion Energy Images (MEI) and Motion History Images (MHI) to recognize many types of aerobics exercises. While their method is efficient, their work assumes that the actor is well segmented from the background and centred in the image.

Schuldt [4] proposed a method for recognizing complex motion patterns based on local space-time features in video and demonstrated such features can give good classification performance. They construct video representations in terms of local space-time features and integrate such representations with SVM classification schemes for recognition.

Ke [5] studies the use of volumetric features as an alternative to the local descriptor approaches for event detection in video sequences. They generalize the notion of 2D box features to 3D spatio-temporal volumetric features. They construct a real-time event detector for each action of interest by learning a cascade of filters based on volumetric features that efficiently scans video sequences in space and time. This event detector recognizes actions that are traditionally problematic for interest point methods such as smooth motions where insufficient space-time interest points are available. Their experiments demonstrate that the technique accurately detects actions on real-world sequences and is robust to changes in viewpoint, scale and action speed.

Weinland [6] introduces Motion History Volumes (MHV) as a free-viewpoint representation for human actions in the case of multiple calibrated, and background-subtracted, video cameras.

We note that the feature vector in these two methods is very expensive to construct and the learning process is difficult, because it needs a big data set for training.

Wong and Cipolla [7] proposed a new method to recognise primitive movements based on the Motion Gradient Orientation (MGO) image directly from image sequences. This process extracts the descriptive motion feature without depending on any tracking algorithms. By using a sparse Bayesian classifier, they obtained good classification results for human gesture recognition.

Ogata [8] proposed another efficient technique for human motion recognition based on motion history images and an eigenspace technique. In the proposed technique, they use Modified Motion History Images (MMHI) feature images and the eigenspace technique to realize high-speed recognition. The experiment results showed satisfactory performance of the technique. However, the eigenspace still needs to be constructed and sometimes this is difficult.

Recently, Dalal [9] proposed a Histogram of Oriented Gradient (HOG) appearance descriptors for image sequences and developed a detector for standing and moving people in video. In this work, several different motion coding schemes were tested and it was shown empirically that orientated histograms of differential optical flow give the best overall performance.

Oikonomopoulos [10] introduced a sparse representation of image sequences as a collection of spatiotemporal events that are localized at points that are salient both in space and time for human actions recognition.

These two methods need to detect salient points in the frames and then make suitable features for classification. This implies significant computational cost for detecting these points.

Meng [11] proposed a fast system for human action recognition which was based on very simple features. They chose MHI, MMHI, MGO and a linear classifier SVM for fast classification. Experimental results showed that this system could achieve good performance in human action recognition. Further, they [12] proposed to combine two kinds of motion features MHI and MMHI together and achieved better performance in human action recognition based on a linear SVM_2K classifier [13] [14]. However, both these systems could only work well in specific real-time applications with limited action classes because the overall performance on real-world challenging database were still not good enough.

3 Overall Architecture

We propose a novel architecture for fast human action recognition. In this architecture, a linear SVM was chosen and MHI provided our fundamental features. In contrast with the system in [11], we propose novel extraction methods to extract both spatial and temporal information from these initial MHI features and combine them in an efficient way as a new feature vector that has lower dimension and provides better motion action information than the raw MHI feature vector.

There are two reasons for choosing a linear SVM as the classifier in the system. Firstly SVM is a classifier that has achieved very good performance in lots of real-world classification problems. Secondly, SVM can deal with very high dimensional feature vectors, which means that there is plenty of freedom to choose the feature vectors. Finally the classifier is able to operate very quickly during the recognition process.

The overall architecture of the human action system is shown in figure 1. There are two parts in this system: a learning part and a classification part.

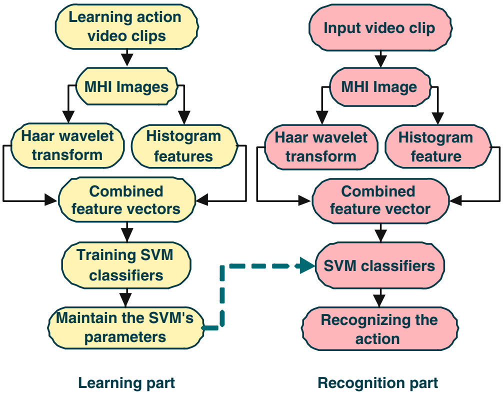


Fig. 1. SVM based human action recognition system. In the learning part, the combined feature vector of Haar wavelet transform and histogram of MHI were used for training a SVM classifier, and the obtained parameters were used in the recognition part.

The MHI feature vectors are obtained directly from human action video clips. The 2-D Haar wavelet transform was employed to extract spatial information within the MHI, while temporal information was extracted by computing the histogram of the MHI. Then these two feature vectors were combined to produce a lower dimensional and discriminative feature vector. Finally, the linear SVM was used for the classification process.

The learning part is processed using video data collected off-line. After that, the obtained parameters for the classifier can be used in a small, embedded computing device such as a field-programmable gate array (FPGA) or digital signal processor (DSP) based system, which can be embedded in the application and give real-time performance.

It should be mentioned here that, both 2-D Haar wavelet transform and histogram of the MHI are achieved with very low computational cost. We only keep the low-frequency part of the Haar wavelet transform. So the total dimension of the combined feature vector is lower than that of the original MHI feature.

4 Detail of the Method

In this section, we will give the detailed information of the key techniques used in our human action recognition system.

4.1 Motion Features

The recording of human actions usually needs large amounts of digital storage space and it is time consuming to browse the whole video to find the required information. It is also difficult to deal with this huge data in detection and recognition. Therefore, several motion features have been proposed to compact the whole motion sequence into one image to represent the motion. The most popular of these are the MHI, MMHI and MGO. These three motion features have the same size as the frame of the video, but they maintain the motion information within them. In [11], it has been found that MHI achieved best performance in classification tests across six categories of action sequence.

A motion history image (MHI) is a kind of temporal template. It is the weighted sum of past successive images and the weights decay as time lapses. Therefore, an MHI image contains past raw images within itself, where most recent image is brighter than past ones.

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

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

where $D(u,v,k)$ is a binary image obtained from subtraction of frames, and τ is the maximum duration a motion is stored. In general, τ is chosen as constant 255 where MHI can be easily represented as a grayscale image. An MHI pixel can have a range of values, whereas the Motion Energy Image (MEI) is its binary version. This can easily be computed by thresholding $H_\tau > 0$.

4.2 Histogram of MHI

The histogram of the MHI has bins which record the frequency at which each value (gray-level) occurs in the MHI, excluding the zero value, which does not contain any motion information of the action. Thus, typically we will have bins between 1 and 255 populated by one or more groupings, where each grouping of bins represents a motion trajectory. Clearly the most recent motion is at the right of the histogram, with the earliest motions recorded in the MHI being more toward the left of the histogram. The spread of each grouping in the histogram indicates the speed of the motion, with narrow groupings indicating fast motions and wide groupings indicating slow motions.

4.3 Haar Wavelet Transform

The Haar wavelet transform decomposes a signal into a time-frequency field based on the Haar wavelet function basis. For discrete digital signals, the discrete wavelet transform can be implemented efficiently by Mallat's fast algorithm [15]. The Mallat algorithm is in fact a classical scheme known in the signal processing community as a two-channel subband coder (see page 1 Wavelets and Filter Banks, by Strang and Nguyen [16]).

The Mallat algorithm is used for both wavelet decomposition and reconstruction. The algorithm has a pyramidal structure with the underlying operations being convolution and decimation. For a discrete signal $s = (s_0, s_1, \cdots, s_{N-1})(N = 2^L, L \in Z^+)$. For convenience, denote it as $c_{m,0} = s_m, m = 0, 1, \cdots, N-1$. Then Haar wavelet transform can be implemented by the following iteration:
For $l = 1, 2, \cdots, L$ and $m = 0, 1, \cdots, N/(l+1)$

$$\begin{cases} c_{m,l} = \sum\limits_{k=0}^{1} h_k c_{k+2m,l-1} \\ d_{m,l} = \sum\limits_{k=0}^{1} g_k c_{k+2m,l-1} \end{cases} \tag{2}$$

where the h_0, h_1 is low pass filter and g_0, g_1 is high pass filter:

$$h_0 = h_1 = \sqrt{2},$$
$$g_0 = \sqrt{2}, g_1 = -\sqrt{2}$$

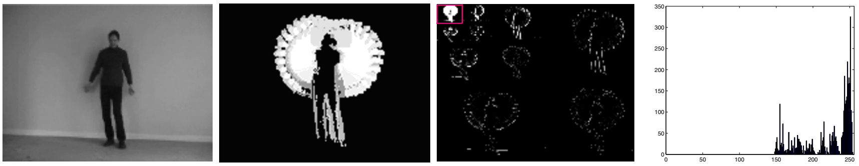
They are orthogonal:

$$g_k = (-1)^k h_{1-k} \tag{3}$$

and the obtained $\{c_{.,L}, d_{.,L}, d_{.,L-1}, \cdots, d_{.,1}\}$ is the discrete Haar wavelet transform of the signal.

An image is a 2-D signal and this 2-D space can be regarded as a separable space, which means that the wavelet transform on an image can be implemented using a 1-D wavelet transform. On the same level, it can be implemented on all the rows and then on all the columns.

In this paper, we only keep the low-frequency part of the Haar wavelet transform of the image. This part can represent the spatial information of the MHI very well in a lower dimension. The high-frequency information is more useful for representing edges, which is not really important in our system, and it is more susceptible to noise. Actually, this part can be implemented very quickly based on some specific algorithms.



(a) Original video (b) MHI (c) Wavelet transform (d) Histogram

Fig. 2. Motion feature of the action. (a) Original video (b)MHI (c) Haar wavelet transform of MHI (d) Histogram of MHI.

Figure 2 shows an example of a handwaving action. (a) is the original video clip, (b) is the MHI, (c) is the Haar wavelet transform of MHI where the red square is low frequency part and (d) is Histogram of the MHI.

4.4 Combining Features

The two feature vectors *histogram of MHI* and *Haar wavelet transform of MHI* are combined in the simplest way. The combined feature vector is built by concatenating these two feature vectors into a higher dimensional vector. In this way, the temporal and spatial information of the MHI are integrated into one feature vector while the dimension of the combined feature vector has lower dimension in comparison with MHI itself.

4.5 Support Vector Machine

SVM is a state-of-the-art classification technique with large application in a range of fields including text classification, face recognition and genomic classification, where patterns can be described by a finite set of characteristic features. We use the SVM for the classification component of our system. This is due to SVM being a classifier that has excellent performance on many real-world classification problems. Using arbitrary positive definite kernels provides a possibility to extend the SVM capability to handle high dimensional feature spaces.

Originally, the SVM is a binary classifier in a higher dimensional space where a maximal separating hyperplane is constructed. Two parallel hyperplanes are constructed on each side of the hyperplane that separates the data. The separating hyperplane is the hyperplane that maximizes the distance between the two parallel hyperplanes. If we have a training dataset $\left\{\mathbf{x}_i|\mathbf{x}_i \in R^d\right\}$, and its binary labels are denoted as $\{y_i|y_i = \pm 1\}$, the norm-2 soft-margin SVM can be represented as a constrained optimization problem

$$\min_{w,b,\xi} \frac{1}{2}||\mathbf{w}||^2 + C\sum_i \xi_i \tag{4}$$

s.t.

$$\begin{gathered}\langle \mathbf{x}_i, \mathbf{w}\rangle + b \geq 1 - \xi_i,\ y_i = 1,\\ \langle \mathbf{x}_i, \mathbf{w}\rangle + b \leq -1 + \xi_i,\ y_i = -1,\\ \xi_i \geq 0,\end{gathered}$$

where C is a penalty parameter and ξ_i are slack variables. The vector $\mathbf{w} \in R^d$ points perpendicular to the separating hyperplane. Adding the offset parameter b allows us to increase the margin. It can be converted by applying Lagrange multipliers into its Wolfe dual problem and can be solved by quadratic programming methods.

The primal optimum solution for weight vector $\mathbf{w}$ can be represented as

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i. \tag{5}$$

where $0 \leq \alpha_i \leq C$. Obviously, $\mathbf{w}$ can be expressed as a linear combination of the support vectors for which $\alpha_i > 0$. For a testing feature vector $\mathbf{x}$, the decision function η and its estimated label h are:

$$h\left(\mathbf{x}\right) = sign\left(\eta\left(\mathbf{x}\right)\right) = sign\left(\langle \mathbf{w}, \mathbf{x}\rangle + b\right). \tag{6}$$

The original optimal hyperplane algorithm was a linear classifier. However, many researchers have created non-linear classifiers by applying a kernel trick [17] and thus the SVM can be generalized to the case where the decision function is a non-linear function of the data.

Multiclass SVMs are usually implemented by combining several two-class SVMs. In each binary SVM, only one class is labelled as ”1” and the others labelled as ”-1”. The one-versus-all method uses a winner-takes-all strategy.

If there are M classes, then the SVM method will construct M binary classifiers by learning. During the testing process, each classifier will get a confidence coefficient $\{\eta_j\left(\mathbf{x}\right)|j=1,2,\cdots,M\}$ and the class k with the maximum confidence coefficient will be assigned to this sample $\mathbf{x}$.

$$h\left(\mathbf{x}\right)=k, \qquad if\ \ \eta_k\left(\mathbf{x}\right)=max_{j=1}^{M}\left(\eta_j\left(\mathbf{x}\right)\right). \tag{7}$$

Our human action recognition problem here is a multi-class classification case. If, for example, we have six classes, then six SVM classifiers are trained based on motion features such as the MHI obtained from human action video clips in a training dataset. For each SVM training, one class is labeled as ”1” and the rest classes are labeled as ”-1”. After the training, each SVM classifier is represented by two parameters $\mathbf{w}$ and b. These parameters will be stored in the internal memory of the FPGA. In the recognition process, one inner product between obtained MHI and $\mathbf{w}$ will be calculated and added to b for each SVM classifier. Then the final predicted label for the action video will go to the class with the maximum one in the computed six values.

5 Experimental Results

5.1 Dataset

For the evaluation, we use a challenging human action recognition database, recorded by Christian Schuldt [4]. It contains six types of human actions (walking, jogging, running, boxing, hand waving and hand clapping) performed several times by 25 subjects in four different scenarios: outdoors (s1), outdoors with scale variation (s2), outdoors with different clothes (s3) and indoors (s4).

This database contains 2391 sequences. All sequences were taken over homogeneous backgrounds with a static camera with 25Hz frame rate. The sequences were down-sampled to the spatial resolution of 160×120 pixels. For all the action sequences, the length of the sequences are vary and the average is four seconds (about 100 frames). To the best of our knowledge, this is the largest video database with sequences of human actions taken over different scenarios. All sequences were divided with respect to the subjects into a training set (8 persons), a validation set (8 persons) and a test set (9 persons). In our experiment, the classifiers were trained on the training set while classification results were obtained on the test set.

Figure 3 showed six types of human actions in the database: walking, jogging, running, boxing, handclapping and handwaving. Row (a) are the original videos, (b) and (c) are the associated MHI and Histogram of MHI features.

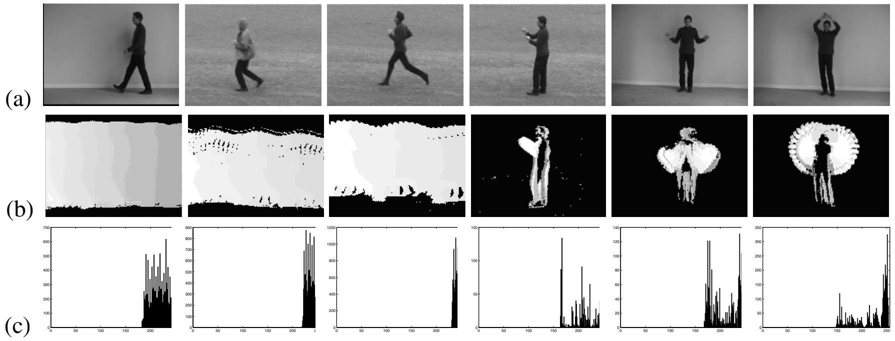


Fig. 3. Six types of human actions in the database: walking, jogging, running, boxing, handclapping and handwaving. Row (a) are the original videos, (b) and (c) are associate MHI and Histogram of MHI features.

5.2 Experimental Setup

Our experiments were carried out on all four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. In the same manner as paper [5], each sequence is treated individually during the training and classification process. In all the following experiments, the parameters were chosen to be the same. The threshold in differential frame computing was chosen as 25 and τ was chosen as constant 255 for MHI construction.

A MHI was calculated from each action sequence with about 100 frames. The size of each MHI is $160 \times 120 = 19200$, which is same width as that of the frames in the videos. The values of MHI are in the interval of $[0, 255]$. Then each MHI was decomposed using a 2-D Haar wavelet transform to $L = 3$ levels. Thus the size of the low frequency part of the Haar wavelet transform of MHI is $20 \times 15 = 300$. Since the length of the histogram of MHI is 255, the length of combined feature vector is 555.

In our system, each SVM was trained based on features obtained from human action video clips in a training dataset. These video clips have their own labels such as ”walking,” ”running” and so on. In classification, we actually get a six-class classification problem. The SVM training can be implemented using programs freely available on the web, such as SVM^{light} [18]. Finally, we obtained several SVM classifiers with associated parameters.

In the recognition process, feature vectors will be extracted from the input human action video sample. Then all the SVM classifiers obtained from the training process will classify the extracted feature vector. Finally, the class with maximum confidence coefficient within these SVM classifiers will be assigned to this sample.

5.3 Experiment Results

Tables 1 show the classification confusion matrix based on the method proposed in paper [5].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

Table 1. Ke's confusion matrix, trace=377.8

	Walk	Jog	Run	Box	Clap	Wave
Walk	**80.6**	11.1	8.3	0.0	0.0	0.0
Jog	30.6	**36.2**	33.3	0.0	0.0	0.0
Run	2.8	25.0	**44.4**	0.0	27.8	0.0
Box	0.0	2.8	11.1	**69.4**	11.1	5.6
Clap	0.0	0.0	5.6	36.1	**55.6**	2.8
Wave	0.0	5.6	0.0	2.8	0.0	**91.7**

Table 2. MHI_S's confusion matrix, trace=377.7

	Walk	Jog	Run	Box	Clap	Wave
Walk	**56.9**	18.1	22.2	0.0	0.0	2.8
Jog	45.1	**29.9**	22.9	1.4	0.0	0.7
Run	34.7	27.8	**36.1**	0.0	0.0	1.4
Box	0.0	0.0	0.0	**89.5**	2.1	8.4
Clap	0.0	0.0	0.0	5.6	**88.9**	5.6
Wave	0.0	0.0	0.0	12.5	11.1	**76.4**

Table 3. MHI_hist's confusion matrix, trace=328.6

	Walk	Jog	Run	Box	Clap	Wave
Walk	**62.5**	32.6	0.0	1.4	1.4	2.1
Jog	12.5	**58.3**	25.0	0.0	0.0	4.2
Run	0.7	18.8	**77.1**	0.0	0.0	3.5
Box	4.9	2.8	0.7	**17.5**	61.5	12.6
Clap	4.9	2.1	0.7	11.1	**75.0**	6.3
Wave	5.6	3.5	6.9	20.1	25.7	**38.2**

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.

In order to study the performance of the Haar wavelet transform of MHI and histogram of MHI, we used linear SVM classifier on them separately and compared their performance. Table 2 and table 3 shows the confusion matrix obtained for Haar wavelet transform and histogram of MHI separately. From these two tables, it can be seen that Haar wavelet transform of MHI obtains a similar performance to Ke's method. This feature did very well in distinguishing the last three groups. On the other hand, histogram of MHI did not do well on overall performance. But it has the power to distinguish the first three groups. That demonstrates that they keep different information from MHI.

Table 4. MHI_S&MHI_hist’s confusion matrix, trace=425.6

	Walk	Jog	Run	Box	Clap	Wave
Walk	**68.8**	11.1	17.4	0.0	0.0	2.8
Jog	36.8	**36.1**	25.0	1.4	0.0	0.7
Run	14.6	20.1	**63.9**	0.0	0.0	1.4
Box	0.0	0.0	0.0	**89.5**	2.1	8.4
Clap	0.0	0.0	0.0	4.9	**89.6**	5.6
Wave	0.0	0.0	0.0	11.1	11.1	**77.8**

Table 4 show the confusion matrix obtained from our system in which combined feature were used. From this table, we can see that the overall performance has got a significant improvement on Ke’s method based on volumetric features. Good performance is achieved in distinguishing all of the six actions in the dataset.

It should be mentioned here that in paper [4], the performance is slightly better where trace=430.3. But our system was trained in the same way as [5] to detect a single instance of each action within arbitrary sequences while Schuldt’s system has the easier task of classifying each complete sequence (containing several repetitions of same action) into one of six classes.

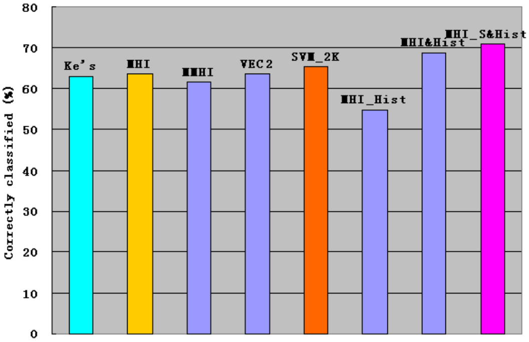


Fig. 4. Comparison results on the correctly classified rate based on different methods: Ke’s method; SVM on MHI; SVM on MMHI; SVM on the concatenated feature (VEC2) of MHI and MMHI and SVM_2K on MHI and MMHI;SVM on histogram of MHI; SVM on the combined feature of MHI and histogram of MHI; SVM on combined feature of Haar wavelet transform of MHI and histogram of MHI.

We also compared the correctly classified rate based on our system with other previous results in the figure 4. The first one is the Ke’s method, the second, third and sixth are SVM based on individual features MHI, MMHI and Histogram of MHI respectively. The fourth one is SVM based on combined feature from MHI and MMHI. The fifth is using SVM_2K classifier on both MHI and MMHI. The seventh is SVM on combined feature from MHI and its histogram. The last one is the results SVM based on Haar wavelet transform of MHI and histogram of MHI. This last result achieves the best overall performance of approximately 71% correct classification.

6 Extension of the Idea

In the previous sections, we combined two different types features extracted from same MHI feature and achieved significant improvement on the performance. The reason is that these two features extract different characteristics of the motion feature. In fact, this idea can be further extended to combine different types of features extracted from different motion features. In [19], we combined the histogram of MHI with Motion Geometric Distribution (MGD) feature vector extracted from the Motion History Histogram. The main common point between MGD feature and Haar wavelet transform feature is that both of them represented spatial information of the motion features. Table 5 showed the experiment results on the same dataset. It achieves the best overall performance of above 80% correct classification.

Table 5. MGD & Hist. of MHI's confusion matrix, trace=481.9

	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**

7 Conclusions

In this paper, we proposed a system for fast human action recognition. Potential applications include security systems, man-machine communication, and ubiquitous vision systems. The proposed method does not rely on accurate tracking as many other works do, since many tracking algorithms incur a prohibitive computational cost for the system. Our system is based on simple features in order to achieve high-speed recognition, particularly in real-time embedded vision applications.

In comparison with local SVM methods by Schuldt [4], our feature vector is much easier to obtain because we don't need to find interest points in each frame. We also don't need a validation dataset for parameter tuning.

In comparison with Meng's [11] [12] methods, we use a Haar wavelet transform and histogram methods to build a new feature vector from the MHI representation. This new feature vector contains the important information of the MHI and also has a lower dimension. Experimental results demonstrate that these techniques made a significant improvement on the human action recognition performance compared to other methods.

If the learning part of the system is conducted off-line, this system has great potential for implementation in small, embedded computing devices, typically FPGA or DSP based systems, which can be embedded in the application and give real-time performance.

Acknowledgements. This work is supported by the DTI (UK) and Broadcom Ltd.

References

1. Aggarwal, J.K., Cai, Q.: Human motion analysis: a review. Comput. Vis. Image Underst. 73, 428–440 (1999)
2. Farnell, B.: Moving bodies, acting selves. Annual Review of Anthropology 28, 341–373 (1999)
3. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. 23, 257–267 (2001)
4. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: Proc. Int. Conf. Pattern Recognition (ICPR 2004), Cambridge, U.K (2004)
5. Ke, Y., Sukthankar, R., Hebert, M.: Efficient visual event detection using volumetric features. In: Proceedings of International Conference on Computer Vision, Beijing, China, October 15-21, pp. 166–173 (2005)
6. Weinland, D., Ronfard, R., Boyer, E.: Motion history volumes for free viewpoint action recognition. In: IEEE International Workshop on modeling People and Human Interaction (PHI 2005) (2005)
7. Wong, S.F., Cipolla, R.: Real-time adaptive hand motion recognition using a sparse bayesian classifier. In: ICCV-HCI, pp. 170–179 (2005)
8. Ogata, T., Tan, J.K., Ishikawa, S.: High-speed human motion recognition based on a motion history image and an eigenspace. IEICE Transactions on Information and Systems E89, 281–289 (2006)
9. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 428–441. Springer, Heidelberg (2006)
10. Oikonomopoulos, A., Patras, I., Pantic, M.: Kernel-based recognition of human actions using spatiotemporal salient points. In: Proceedings of IEEE Int'l Conf. on Computer Vision and Pattern Recognition 2006, vol. 3 (2006)
11. Meng, H., Pears, N., Bailey, C.: Recognizing human actions based on motion information and svm. In: 2nd IET International Conference on Intelligent Environments, Athens, Greece, IET, pp. 239–245 (2006)
12. Meng, H., Pears, N., Bailey, C.: Human action classification using svm_2k classifier on motion features. In: Gunsel, B., Jain, A.K., Tekalp, A.M., Sankur, B. (eds.) MRCS 2006. LNCS, vol. 4105, pp. 458–465. Springer, Heidelberg (2006)
13. Meng, H., Shawe-Taylor, J., Szedmak, S., Farquhar, J.D.R.: Support vector machine to synthesise kernels. In: Deterministic and Statistical Methods in Machine Learning, 242–255 (2004)
14. Farquhar, J.D.R., Hardoon, D.R., Meng, H., Shawe-Taylor, J., Szedmak, S.: Two view learning: Svm-2k, theory and practice. In: NIPS (2005)
15. Mallat, S.: A theory for multiresolution signal decomposition: the wavelet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 11, 674–693 (1989)
16. Strang, G., Nguyen, T.: Wavelets and Filter Banks. Wellesley Cambridge Press (1996)
17. Aizerman, A., Braverman, E.M., Rozoner, L.I.: Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control 25, 821–837 (1964)
18. Joachims, T.: Making large-scale svm learning practical. In: Oikonomopoulos, A., Patras, I., Pantic, M. (eds.) Advances in Kernel Methods - Support Vector Learning, USA. MIT-Press, Cambridge (1999)
19. Meng, H., Pears, N., Bailey, C.: A human action recognition system for embedded computer vision application. In: The 3rd IEEE workshop on Embeded Computer Vision, Minneapolis,USA (2007)