# Recognition of human activities using a multiclass relevance vector machine

He, Weihua; Yow, Kin Choong; Guo, Yongcai

2012

He, W., Yow, K. C., & Guo, Y. (2012). Recognition of human activities using a multiclass relevance vector machine. Optical Engineering, 51(1).

https://hdl.handle.net/10356/98856

https://doi.org/10.1117/1.OE.51.1.017202

Optical Engineering

SPIEDigitalLibrary.org/oe

# Recognition of human activities using a multiclass relevance vector machine

Weihua He
Kin Choong Yow
Yongcai Guo

# Recognition of human activities using a multiclass relevance vector machine

**Weihua He**
Chongqing University
Key Laboratory of Optoelectronic Technology and Systems
Ministry of Education
Chongqing 400030, China
and
Nanyang Technological University
School of Computer Engineering
N4-B2c-06, Nanyang Avenue
Singapore 639798, Singapore
E-mail: heweihua@cqu.edu.cn

**Kin Choong Yow**
Nanyang Technological University
School of Computer Engineering
N4-B2c-06, Nanyang Avenue
Singapore 639798, Singapore

**Yongcai Guo**
Chongqing University
Key Laboratory of Optoelectronic Technology and Systems
Ministry of Education
Chongqing 400030, China

**Abstract.** We address the issue of human activity recognition by introducing the multiclass relevance vector machine (mRVM), the current state-of-the-art kernel machine learning technology given the multiclass classification problems (actually, activity recognition can commonly be viewed as a multiclass classification problem). Under our proposed recognition framework, the required procedure consists of three functional cascade modules: a. detecting the human silhouette blobs from the image sequence by the background subtraction method; b. extracting the shape and the motion features from the variation energy image (VEI); and c. sending the obtained features to the mRVM and recognizing the human activity. There are two types of mRVM: the constructive $mRVM_1$ and the top-down $mRVM_2$. We performed 10 times three-fold cross-validation on the Weizmann benchmark data set to examine the effectiveness of the proposed method. We also compared our method with other existing approaches, and the experimental results show that the proposed method offers superior performance. In summary, the mRVM, especially the $mRVM_2$, has advantages both in terms of recognition rate and sparsity, along with a simple feature extraction process. The mRVM also significantly simplifies the classification process, by comparison with traditional binary-tree style multiclass classifiers.  [DOI: 10.1117/1.OE.51.1.017202]

Subject terms: human activity recognition; kernel machine learning; multiclass relevance vector machine; variation energy image.

Paper 110495 received May 6, 2011; revised manuscript received Oct. 2, 2011; accepted for publication Nov. 8, 2011; published online Feb. 6, 2012.

## 1 Introduction

Human activity recognition has received a great amount of attention in the computer vision and video processing communities. Research in this domain is motivated by a wide range of promising applications, such as visual surveillance,[1] analysis of sports events,[2] human-computer interaction,[3] and biometrics.[4] The procedure of human activity recognition includes two stages: a. human detection and activity features extraction, i.e., representation of activity, and b. classification of activity. The selection of methodologies used in both stages has a great influence on the final recognition performance.

In the past few years a lot of work has been done on human activity and motion analysis to search for a reasonable representation of human activity. The existing methods can be classified into two categories: model-based methods and model-free methods. The model-based methods have been widely used in motion analysis and the relevant applications, such as gait recognition. Many two-dimensional and three-dimensional human models have been proposed recently. Rectangle and skeleton models[5,6] are the most frequently adopted models in this category. Due to the large degree of freedom in the non-rigid human body and the high variability of individual shape, the model-based approaches usually lack generality; thus, their applicability to actual situations may be limited. On the other hand, in the model-free methods, human movement is usually represented by low-level features extracted from human appearance or motion. Such methods exhibit much flexibility and robustness and are promising for implementation in real-time applications. Bobick and Davis[7] proposed the motion energy image (MEI) and motion-history image (MHI) for human movement representation and recognition, which were constructed from cumulative binary motion images. Wang[8] introduced the average motion energy (AME) and mean motion shape for human action recognition. Here, we propose a new spatio-temporal template: the variation energy image (VEI), derived from the sequence of the detected silhouette blobs and the corresponding average motion image (AME). Then, we extracted both the shape and the motion features from the VEI, which are used to represent the activity.

The selection of an appropriate classification method also plays a crucial role in a generic human activity recognition system. So far, many well-known pattern recognition techniques, such as $k$-nearest neighbor ($k$-NN), support vector machine (SVM),[4,9] and relevance vector machine (RVM), as well as their variants have been developed and applied in activity recognition. The $k$-NN simply selects $k$-closest samples from the training data, and the class with the highest number of votes is assigned to the observed sequence. However, in $k$-NN the features extracted from all the training samples are employed without selection, which usually means redundancy and thus leads to vain computation cost. In addition, the $k$-NN essentially gives equal weight to each training sample, which may not be appropriate in some cases. The SVM overcomes this problem by selecting the

0091-3286/2012/$25.00 © 2012 SPIE

informative samples located at the boundary of the decision function. Accordingly, only a fraction of samples in the training set are employed for the subsequent classification process, i.e., the SVM can maintain relative sparsity of support vectors without any loss of recognition accuracy. However, the SVM still has some inherent drawbacks: a. the number of support vectors typically grows linearly with the size of the training set, i.e., the computational complexity would increase proportionally to the scale of the problem, and b. the traditional SVM makes non-probabilistic but hard binary decisions. However, the probabilistic prediction is particularly crucial in classifications where the posterior probabilities of class membership are necessary to adapt to the varying class priors and the asymmetric misclassification costs. Some post-processing techniques have been employed to transform the binary outputs to probabilistic outputs for SVMs.[10,11] Nevertheless, Tipping[12] argued that these estimates are unreliable, and the kernel function of the SVM should meet some special requirements. To further overcome the shortcoming of SVM, the relevance vector machine (RVM) was put forward, which relies on Bayesian inference learning. The RVM can achieve comparable or even better performance than the SVM while at the same time the sparsity of relevance vectors is greatly improved. Gholami et al.[4] distinguished pain from non-pain in neonates and assessed their pain intensity by using the RVM classification technique. Selvathi[9] used RVM and SVM for the classification of volume of MRI data as normal or abnormal and demonstrated that the RVM can obtain higher classification accuracy than SVM.

Since both RVM and SVM were originally devised for a two-class context, they must be modified so as to be applicable in the multiclass discrimination problems. One possible solution is to construct multiple RVM classifiers and combine them together. For instance, B. Yogameena et al.[13] extended the RVM to a multiclass scheme by training multiple mapping functions to implement the sub-classification; then, these sub-classification results were combined for the final pose recognition. Since this method needs to train multiple RVM classifiers, the advantage of sparsity that SVM or RVM has would be more or less canceled out. Another possible solution is to directly generalize the RVM to the multiclass RVM. But it had been viewed as impractical for some time due to the bad scaling of the type-II ML procedure with respect to the number of classes $C$[14] and the dimensionality of the Hessian required for the Laplace approximation.[15] Until very recently, a novel classification algorithm, the multiclass relevance vector machine (mRVM), has been introduced, which expands the original RVM to the multiclass and multi-kernel setting.[15] Psorakis et al.[16] provided further insight into the theoretical background of the mRVM and proposed a collection of methodologies that boost the performance of mRVM, both in terms of computational complexity and discrimination power. The mRVM gains multiclass discrimination by introducing an auxiliary variable $Y$, from which we can derive the multinomial probit likelihood for the estimation of class membership probabilities. Different from the method proposed in the reference,[13] the mRVM does not need the extra procedure to minimize the cost functions of multiple RVM regression. Moreover, it does not need to train multiple RVM classifiers but can use merely a single mRVM classifier when dealing with multiclass problems. Therefore, the recognition systems are expected to be more efficient and stable if the mRVM is integrated into them.

In this paper we use the mRVM technique[15,16] to recognize human activities. To the best of our knowledge, we are the first to use mRVM to recognize human activities. Figure 1 shows the overview of the proposed activity recognition method. First, the binary silhouettes are detected from the captured image sequences by using the background subtraction technique. After normalization, these binary silhouettes are used to determine the motion period. Subsequently, the VEI is generated, and the activity features are extracted from it. Finally, the extracted features are used by the multiclass classifier mRVM to learn and recognize the human activities.

The rest of the paper is organized as follows: in Sec. 2 we briefly cover the related literature. Section 3 describes the extraction of the human blobs by background subtraction and the VEI construction and representation. Section 4 reviews two different structures of mRVM classification techniques: $\text{mRVM}_1$ and $\text{mRVM}_2$. Then, Sec. 5 reports the training of the classifiers as well as the recognition results on a benchmark database. The last section presents our conclusions and discusses future work.

## 2 Related Work

A number of approaches for human activity recognition have been proposed recently. A detailed overview of these vision-based human activity recognition algorithms can be found in Ref. 17. It covers much work about image representation and activity recognition. Here, we will concentrate mainly on spatio-temporal templates-based activity recognition algorithms, which are more relevant to our work.

Bobick and Davis[7] pioneered the idea of spatio-temporal templates. They characterized each action with a binary motion energy image (MEI) and a motion-history image (MHI), from which seven Hu moments were extracted. Then, the Hu moments were matched using a nearest neighbor (NN) approach based on the Mahalanobis distance in order to recognize the activities. Ahad[18] extended this work by using the directional motion-history image (DMHI) and the NN approach for action recognition.

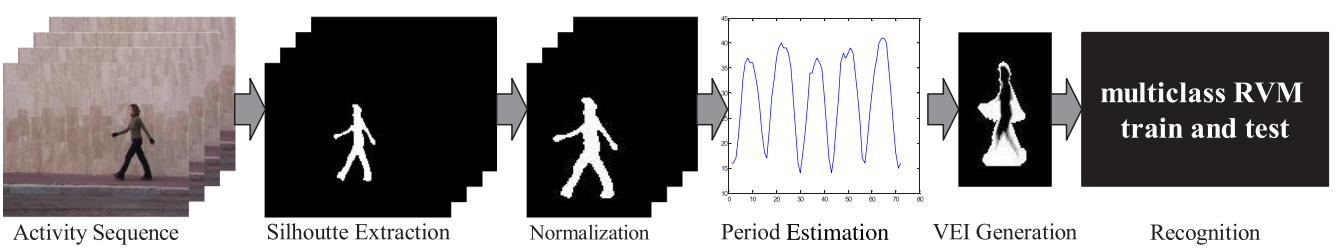


**Fig. 1** Flow chart of the proposed activity recognition method.

Schuldt et al.[19] proposed a method for recognizing complex motion patterns based on local space-time features in video. They integrated these features with an SVM scheme for recognition.

Weinland et al.[20] 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.

Wang et al.[8] introduced average motion energy (AME) and mean motion shape (MMS) to depict human motion in a two-dimensional space. The features extracted from AME and MMS were used to train NN and *k*-NN for recognizing a set, including nine types of natural actions.

Meng et al.[21] proposed a new motion-history histogram (MHH) feature to capture information in a video. To further improve classification performance, they combined both MHI and MHH into a low-dimensional feature vector, which is then processed by a support vector machine (SVM).

Qian et al.[22] introduced a contour coding of the motion energy image (CCMEI) and recognized the activities of people with an SVM multiclass classifier with binary-tree architecture, which is determined by a clustering process.

We note that, in some of these methods, the motion features employed are relatively complex,[20] or the classifiers do not have good discriminative performance,[7,8] which would hinder their feasibility in real applications.

## 3 Feature Extraction of Activities

The original pixel-level data contained in the captured image sequences correspond to extremely high dimensionality and have a lot of redundancy. The dimensionality of the data needs to be reduced through an efficient and effective feature extraction process to ensure real-time activity recognition. Moreover, the reliable detection of a silhouette is also a vital pretreatment for the subsequent feature extraction.

In this section we will briefly introduce the silhouette detection and the features extraction procedures. We use a new spatio-temporal template, VEI, for the feature extraction.

### 3.1 *Silhouette Detection and Post-Processing*

A silhouette is usually detected by finding the difference between the background and the current image[23] or by grouping optical flow to find coherent motion.[24] In our case, we use the background subtraction technique to detect the binary human silhouette. There are inevitably existing spurious pixels, holes inside the moving subject, and other anomalies in the detected sections. The connectivity component analysis and the morphological operations, such as erosion and dilation, are successively applied to remove the spurious pixels and to fill small holes inside the extracted silhouettes. To eliminate the effect of variation in imaging distances, the obtained binary silhouettes are centered and normalized to a fixed size $100 \times 80$. The whole procedure of detecting a human blob is shown in Fig. 2.

### 3.2 *Period or Duration Estimation*

Many types of human activities (e.g., walking, running, jogging, etc.) show periodicity at a short time-duration. The motion period can be determined by calculating the number of foreground pixels in a silhouette image.[25] However, the number of foreground pixels is prone to be affected by noise arisen from complex background, illumination variation, and the change in an object's appearance. Here, we estimated the motion period by examining the aspect ratio values associated with the silhouette image sequence. For example, the silhouette image sequence for a short-time-duration is shown in Fig. 3, which is related to the "walking" activity. Figure 4 shows the corresponding variation in the aspect ratio over the whole time span (here, the whole image sequence includes about five motion cycles). It is clear that the variation in the aspect ratio exhibits strong periodicity; thus, the time duration between two adjacent valley points can be taken as one motion period.

In this paper we assume that all the captured image sequences span more than one motion cycle. Once the motion period parameter is determined, the whole sequences will be truncated into several cycles. Thereafter, the binary silhouettes corresponding to a certain motion cycle are picked out to generate the VEI.


**Fig. 2** Detection of a human silhouette: (a) original image; (b) background image; (c) raw binary silhouette; (d) final silhouette.

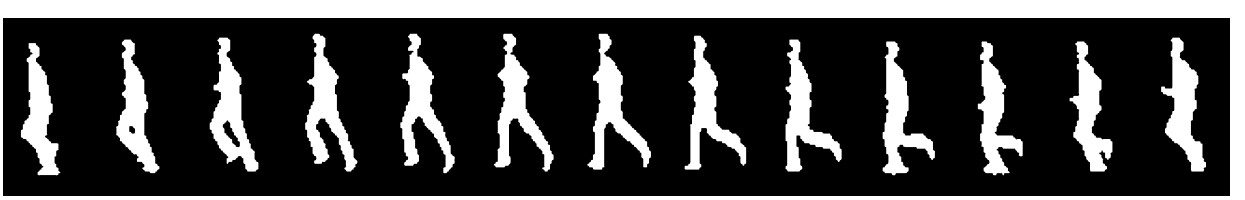

**Fig. 3** Silhoutte sequences in a motion cycle.

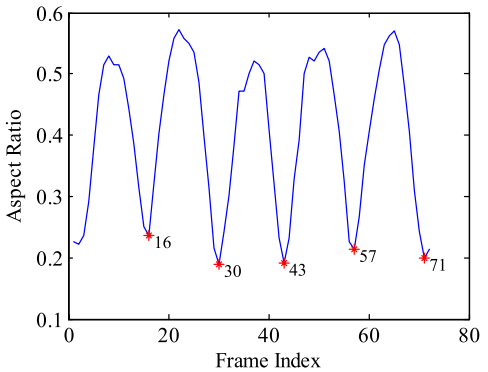


**Fig. 4** Variation of the aspect ratio over time.

### 3.3 *A New Spatio-Temporal Template: VEI*

The spatio-temporal templates-based activity recognition methods try to map the whole captured image sequences into a single or several images,[7,18] which in essence is a data compression process. Here, we proposed a new spatio-temporal template named VEI, which has been extended from the average motion energy (AME).[8] The AME is generated from silhouette sequences in a certain motion cycle as:

$$U(x,y) = \frac{1}{T}\sum_{k=k_0}^{k_0+T-1} f(x,y,k), \tag{1}$$

where $f(x,y,k)$ is the silhouette image corresponding to the $k$-th frame while $T$ equals the number of frames included in a motion cycle. We can also obtain the deviation image by:

$$\sigma(x,y) = \sqrt{\frac{1}{T-1}\sum_{k=k_0}^{k_0+T-1}[f(x,y,k) - U(x,y)]^2}. \tag{2}$$

By using the AME and the deviation image, we compute the VEI as:

$$V(x,y) = \frac{\sigma(x,y)}{U(x,y)}. \tag{3}$$

Figure 5 shows several silhouette samples related to the "jacking" activity, as well as the resultant VEI. Both the shape profile and the motion characteristics are reflected in the VEI.

Next, we will extract the shape and the motion features from the VEI, which would then be integrated and sent to a classifier for training and testing.

#### 3.3.1 *Shape features*

A specific activity is usually accompanied with unique variations in shape profile. For instance, the "bend" activity would greatly expand the body in the lateral direction while the "walking" activity creates variation of shape width within a relative small range, so the aspect ratio metric would discriminate between the "bend" and the "walking" activity. Meanwhile, there also exist significant differences in the centroid positions in such situations even if the same persons are involved.

Based on the aforementioned discussions, we decided to adopt three kinds of shape features for activity recognition: the centroid of AME, the centroid of VEI, and the aspect ratio of VEI.

To find the centroid of the VEI, first we compute its 0-rank and 1-rank geometric moments:

$$\begin{aligned} m_{00} &= \sum_{x=1}^{M}\sum_{y=1}^{N} V(x,y), \\ m_{10} &= \sum_{x=1}^{M}\sum_{y=1}^{N} x \bullet V(x,y), \\ m_{01} &= \sum_{x=1}^{M}\sum_{y=1}^{N} y \bullet V(x,y), \end{aligned} \tag{4}$$

where $V(x,y)$ represents the gray scale value of the VEI at the pixel $(x,y)$ while $M$ and $N$ are the width and the height of the VEI, respectively. Then, the centroid of VEI is defined as

$$(x_0, y_0) = \left(\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}}\right). \tag{5}$$

The centroid $(x_1, y_1)$ of AME can be calculated in a similar way while the aspect ratio of the VEI can be determined from the bounding box surrounding it.

Certainly, the shape of VEI can also be described in some other more deliberate and complicated ways. For instance, we can: a. extract the Hu moments or the Zernike moments[18] from the VEI, which are invariant to the rotation, scale, or translation transforms in the same plane or b. use curve-fitting methods, such as the B-spline method, to describe the contour of VEI, so the fitting parameters will be included in the feature data or c. even do spatial/transformed domain analysis on the landpoints located at the boundary of VEI, such as the square-to-circular transformation,[22] the LLE[26] method, the Fourier transform, etc. By using such shape features during the classification, we may expect better recognition performance. However, the extraction of more

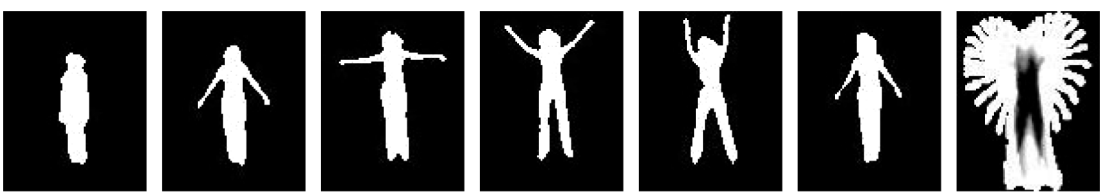


**Fig. 5** Human activity representation using VEI: (a) to (f) represent some key frames of one activity (jacking) while (g) shows the resultant VEI.

deliberate features will increase the computational complexity and the storage cost, which make it infeasible for real-time systems. To maintain the real-time performance of the proposed algorithm, we suggest using relatively simple shape or motion features.

#### 3.3.2 *Motion features*

As shown in Fig. 5(g), the motion characteristics are strengthened in the VEI. More importantly, the VEI takes on a distinctive pattern, from which the corresponding style of activity can be identified. Here, we have extracted three kinds of motion features: the normalized effective area of motion domain, the distribution of motion area in the four quadrants, and the normalized amplitude of the leg raising.

The effective area of motion domain is defined as:

$$S = \sum_{x=1}^{M} \sum_{y=1}^{N} V(x, y). \tag{6}$$

We normalize the effective area of the motion domain ($\bar{S}$) to compensate for the inconsistency among the body profiles of different persons. Although we can measure the motion component from $\bar{S}$ quantitatively, we cannot extract any information about where the motion had taken place. Therefore, the distribution of motion area in the four quadrants is an indispensible and supplementary feature.

We also noticed that the motion of legs always plays an important role in many types of human activity, such as "walking," "skipping," running," etc. Hence, the cue of leg motion cannot be ignored in real applications. For simplicity, we considered only the amplitude of the leg in this paper, which can easily distinguish between the "walking" activity and the "running" activity. The amplitude of the leg raising can be easily determined from the contour of VEI. We normalized it to the body height to eliminate the effect of individual discrepancy.

In summary, six features determined by 11 parameters were extracted from the VEI template. Since our method is based on motion periodicity, the frames that belong to a single motion cycle are required for feature extraction, which reduces the computation cost significantly. Both the shape features and the motion features are used to depict the activities. Therefore, the limitation resulting from either shape or motion-based representation can be alleviated. Moreover, all these adopted features are simple, intuitive, and relatively easy to extract. Thus, they are especially applicable for real-time tasks.

## 4 Multi-Kernel Relevance Vector Machines

In this section we solve the multiclass data classification problem by using a new state-of-the-art sparse Bayesian learning approach, which is known as mRVM. The mRVM expands the original RVM to the multiclass setting by introducing auxiliary variable $Y$, which acts as the intermediate regressing target and can lead to the multinomial probit likelihood for the estimation of class membership probabilities.[16] Based on the search of the relevance vectors during the training phase, two variants ($\text{mRVM}_1$ and $\text{mRVM}_2$) of mRVM[15,16] have been proposed recently. We use both of these variants to classify the human activities.

### 4.1 *The $\text{mRVM}_1$ Classifier*

Assume a training set $\{X_1, X_2, \cdots, X_N\} \subset \mathbb{R}^D$, with target values given by $\{t_1, t_2, \cdots, t_N\}$, where $X_n$ is a $D^s$-dimensional sample and $t_n \in \{1, \cdots, C\}, n = 1, \cdots, N$. The corresponding multinomial likelihood function is as follows:

$$p(t_n = i|\mathbf{W}, \mathbf{k}_n) = \int p(t_n = i|\mathbf{y}_n) p(\mathbf{y}_n|\mathbf{W}, \mathbf{k}_n) \mathrm{d}\mathbf{y}_n = \varepsilon_{p(u)} \left\{ \prod_{j \neq i} \Phi[u + (\mathbf{w}_i - \mathbf{w}_j)]^T \mathbf{k}_n \right\}, \tag{7}$$

where $Y$ is the auxiliary variable introduced for multiple class discrimination and assumed to obey a standardized noise model $y_{nc}|\mathbf{w}_c, \mathbf{k}_n^\beta \sim N_{y_{cn}}(\mathbf{w}_c \mathbf{k}_n^\beta, 1)$, where $\mathbf{w}_c$ from $\mathbf{W}$ follows a standard zero-mean Gaussian distribution $\omega_{nc} \sim N[0, (1/\alpha_{nc})]$, and each element $\mathbf{k}_n$ from the kernel $\mathbf{K}$ describes a similarity measure between activity samples based on specific features.

Given $\mathbf{A} = (\alpha_1, \cdots, \alpha_n)$, the posterior parameter distribution conditioned on the data is given by combining the likelihood with the prior under the Bayes rule:

$$p(\mathbf{W}|\mathbf{Y}, \mathbf{A}, \mathbf{K}) = p(\mathbf{Y}|\mathbf{W}, \mathbf{K}) p(\mathbf{W}|\mathbf{A}) / p(\mathbf{Y}|\mathbf{K}, \mathbf{A}). \tag{8}$$

The value of $\mathbf{A}^*$ can be determined by maximizing the multiclass marginal likelihood $p(\mathbf{Y}|\mathbf{K}, \mathbf{A})$, where

$$p(\mathbf{Y}|\mathbf{K}, \mathbf{A}) = \int p(\mathbf{Y}|\mathbf{K}, \mathbf{W}) p(\mathbf{W}|\mathbf{A}) \mathbf{dW}.$$

The logarithm counterpart $L(A)$ can be further expanded as:

$$L(\mathbf{A}) = \log p(\mathbf{Y}|\mathbf{K}, \mathbf{A}) = \log \int p(\mathbf{Y}|\mathbf{K}, \mathbf{W}) p(\mathbf{W}|\mathbf{A}) \mathbf{dW} = \sum_{c=1}^{C} -\frac{1}{2} [N \log 2\pi + \log |\mathbf{C}| + \mathbf{y}_c^T \mathbf{C}^{-1} \mathbf{y}_c], \tag{9}$$

where $\mathbf{C} = \mathbf{I} + \mathbf{K}\mathbf{A}^{-1}\mathbf{K}^T$ and the determinant and inverse of $\mathbf{C}$ can then be written as:

$$|\mathbf{C}| = |\mathbf{C}_{-i}||1 + \alpha_i^{-1} k_i^T \mathbf{C}_{-i}^{-1} k_i|,$$

and

$$\mathbf{C}^{-1} = \mathbf{C}_{-i}^{-1} - \frac{\mathbf{C}_{-i}^{-1} k_i k_i^T \mathbf{C}_{-i}^{-1}}{\alpha_i + k_i^T \mathbf{C}_{-i}^{-1} k_i}.$$

We can then rewrite the logarithm marginal likelihood function [Eq. (11)] in the form

$$L(\mathbf{A}) = L(\mathbf{A}_{-i}) + \sum_{c=1}^{C} \frac{1}{2} \left[ \log \alpha_i - \log(\alpha_i + s_i) + \frac{q_{ci}^2}{\alpha_i + s_i} \right], \tag{10}$$

where $s_i = \mathbf{k}_i^T \mathbf{C}_{-i}^{-1} \mathbf{k}_i$ and $q_{ci} = \mathbf{k}_i^T \mathbf{C}_{-i}^{-1} \mathbf{y}_c$. The sparsity factor $s_i$ defines the measure of overlap between a sample $\mathbf{k}_i$ and the ones already included in the model. $q_{ci}$ measures how good the sample is in helping to describe a specific class. Setting the derivation of $\partial L(\mathbf{A})/\partial \alpha_i$ equals to zero, we can obtain the stationary points:

$$\alpha_i = \frac{Cs_i^2}{\sum_{c=1}^{C} q_{ci} + Cs_i^2} \quad \text{if} \quad \sum_{c=1}^{C} q_{ci}^2 > s_i, \tag{11}$$

$$\alpha_i = \infty \quad \text{if} \quad \sum_{c=1}^{C} q_{ci}^2 \leq s_i. \tag{12}$$

Note that the quantity $\theta_i = \sum_{c=1}^{C} q_{ci}^2 - s_i$ represents the contribution of the $i$'th sample to marginal likelihood, so we can set the rule of inclusion or exclusion sample based on the values of $\theta_i$ and $\alpha_i$:

a. IF $\theta_i > 0$ and $\alpha_i < \infty$, THEN update $\alpha_i$ from Eq. (11)
b. IF $\theta_i > 0$ and $\alpha_i < \infty$, THEN set $\alpha_i$ from Eq. (11)
c. IF $\theta_i \leq 0$ and $\alpha_i < \infty$, THEN set $\alpha_i = \infty$ from Eq. (12)

Then, the M-steps and E-steps are used to estimate $\mathbf{W}$ and the posterior expectations of $\mathbf{Y}$ respectively.

$$\hat{\mathbf{w}}_c = (\mathbf{K}\mathbf{K}^T + \mathbf{A}_c)^{-1}\mathbf{K}\mathbf{y}_c^T, \tag{13}$$

$$\tilde{y}_{cn} \leftarrow \hat{\mathbf{w}}_c^T \mathbf{k}_n - \frac{\varepsilon_{p(u)}\{\mathbf{N}_u(\hat{\mathbf{w}}_c^T\mathbf{k}_n - \hat{\mathbf{w}}_i^T\mathbf{k}_n, 1)\Phi_u^{n,i,c}\}}{\varepsilon_{p(u)}\{\Phi_u(u + \hat{\mathbf{w}}_i^T\mathbf{k}_n - \hat{\mathbf{w}}_c^T\mathbf{k}_n)\Phi_u^{n,i,c}\}}, \tag{14}$$

and for the $i$'th class

$$\tilde{y}_{in} \leftarrow \hat{\mathbf{w}}_i^T \mathbf{k}_n - \left(\sum_{j \neq i} \tilde{y}_{jn} - \hat{\mathbf{w}}_j^T \mathbf{k}_n\right), \tag{15}$$

where $\tilde{y}$ denotes the expectation value.

The training procedure involves consecutive updates of the model parameters. The hyper-parameters $\mathbf{A}$, $\mathbf{W}$, and $\mathbf{Y}$ are all updated during each iteration, and the process is repeated until the change in the hyper-parameter values is minimal or the preset number of iterations has been reached. A detailed description of the training process of the $\text{mRVM}_1$ classifier can be found in Ref. 16.

In the classification phase, the test sample $X$ is classified to the class for which the auxiliary variables $y_{in}$, $1 \leq i \leq C$ is maximized:

$$t_n = \arg\max_i (y_{in}) \tag{16}$$

### 4.2 *The $\text{mRVM}_2$ Classifier*

The $\text{mRVM}_2$ follows a top-down strategy by loading the whole training kernel into memory and iteratively removing non-relevant samples. The training phase related to $\text{mRVM}_2$ is similar to the one for $\text{mRVM}_1$, which also requires consecutive updates of the parameters $\mathbf{W}$ and $\mathbf{Y}$ through Eqs. (13)–(15). The only difference lies in that the $\text{mRVM}_2$ does not adopt the marginal likelihood maximization as $\text{mRVM}_1$ but rather employs an extra E-step for the updates of the hyper-parameters $\mathbf{A}$. The selection rule of training samples and other details can be found in Ref. 16.

## 5 Experimental Results

In this section, we use the classification techniques described in Sec. 4 to recognize human activities. We chose the Weizmann data set, which is a publicly available data set well designed for activity recognition. Here, we will offer a brief introduction about the Weizmann data set and then provide the experimental results. The quantitative comparative evaluation between our method and some other methods was made in two ways: on the one hand, both the feature representation and the classification techniques are individually taken out as independent modules for recognition performance evaluation, so as to thoroughly examine the validity of our method. On the other hand, the performance with different methods is also compared directly.

### 5.1 *The Weizmann Data Set*

This data set has been introduced by Blank et al.[27] It consists of 93 uncompressed videos in avi format with low resolution ($180 \times 144$ pixels, 25 frames/s), including 10 different activities (the numbers in parentheses signify the respective number of videos): bending (9), jumping-jack (9), jumping forward on two legs (9), jumping in place on two legs (9), running (10), galloping sideways (9), skipping (10), walking (10), one-hand waving (9), and two-hands waving (9). The videos are colorful, and the actors are of both genders. Some sample frames from the database are shown in Fig. 6, and the resultant VEIs with respect to each style of activity are shown in Fig. 7.

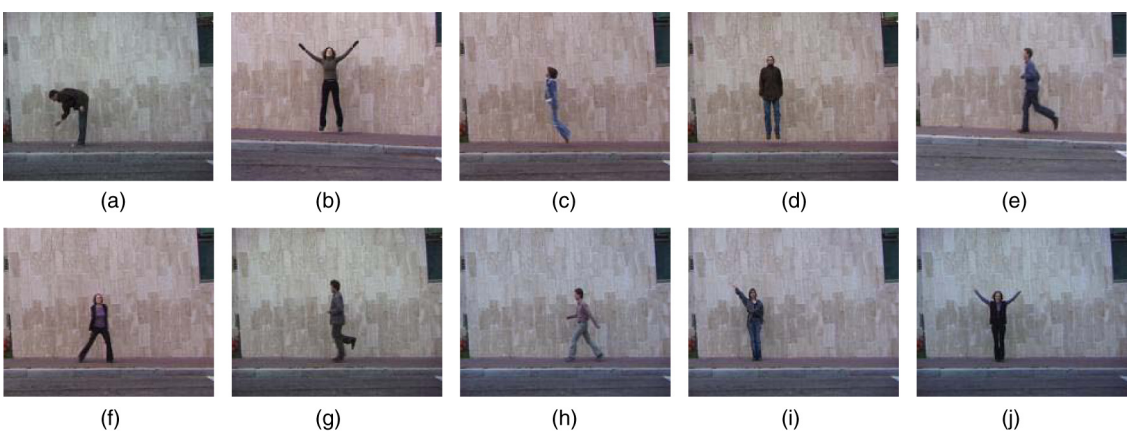


**Fig. 6** Sample frames from the 10 activities in the Weizmann data set: (a) bend, (b) jack, (c) jump, (d) p-jump, (e) run, (f) side, (g) skip, (h) walk, (i) wave 1, (j) wave 2.

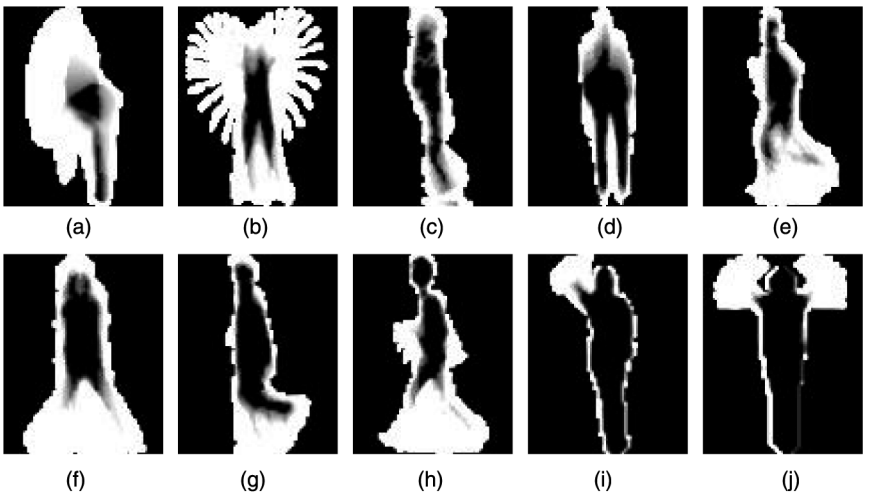


**Fig. 7** VEIs for the ten activities: (a) bend, (b) jack, (c) jump, (d) p-jump, (e) run, (f) side, (g) skip, (h) walk, (i) wave 1, (j) wave 2.

### 5.2 *Activity Recognition with mRVM$_1$ and mRVM$_2$*

As depicted in Sec. 3, the normalized silhouette sequence is used to generate the VEI, from which the activity features are extracted. Next, we will use the mRVM$_1$ and the mRVM$_2$ discussed in Sec. 4 to finish the recognition task.

#### 5.2.1 *Results with the mRVM$_1$*

The mRVM$_1$ is applied to classify the activity features. We perform a 10 times three-fold cross-validation procedure in order to minimize any result variance produced by improper folds. The bandwidth of the Gaussian kernel is an important parameter, which has great influence on the recognition performance. A simple method was used to determine the optimal bandwidth. We increased its value with a constant step of 0.05 over the range of [0.05, 1] and trained the mRVM$_1$ classifier over the whole training set. The bandwidth maximizing the classification accuracy was chosen for the following experiments. Figure 8 shows the training performance of the mRVM$_1$ with different bandwidths, from which we can see the best bandwidth parameter is 0.25. Certainly, the gradient descent methods or other similar optimization methods with variable step size may be used to further improve the classification performance, and this will be considered in our future work.

Figure 9 shows the recognition results with the mRVM$_1$ in terms of percentage accuracy, which includes the worst, the mean, and the best recognition rates of the 10 times validations. The horizontal axis specifies the iteration numbers while the vertical coordinate represents the recognition rates. It can be seen that all the recognition rates initially increase rapidly as the iteration number grows, and after 60 iterations the mean recognition rate gradually becomes constant. A noticeable fact is that the recognition rates witness a heavy fluctuation during the first five iterations, which can be interpreted as a data set-dependent phenomenon. After 120 iterations, the worst, the mean, and the best recognition rates are 89.28%, 94.5%, and 98.19%, respectively. The difference between the worst and the best recognition rates is even as high as 8.91%, which means the train/test split would have great influence on the recognition performance. This phenomenon also suggests that the recognition might be more robust if we can find the optimal training set through multi-times cross-validation.

Figure 10 shows the number of retained relevance vectors (RVs) associated with given iteration numbers, including the minimal, the mean, and the maximal number of RVs in the 10 times validations. Since the mRVM$_1$ is based on the constructive strategy, the RVs start with a single sample and then new RVs are gradually accepted or abandoned as the

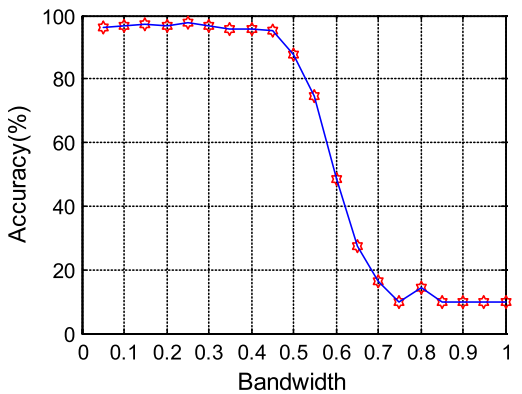


**Fig. 8** The training performance of mRVM$_1$ with different bandwidths.

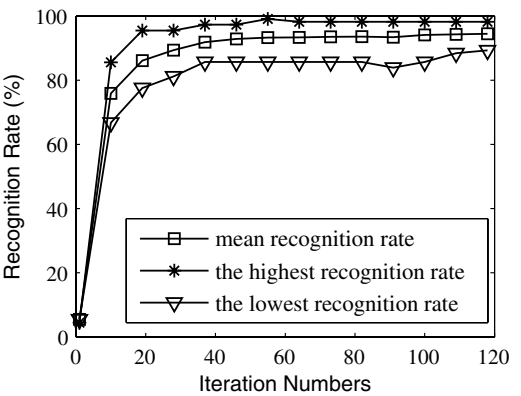


**Fig. 9** Recognition rate with mRVM$_1$.

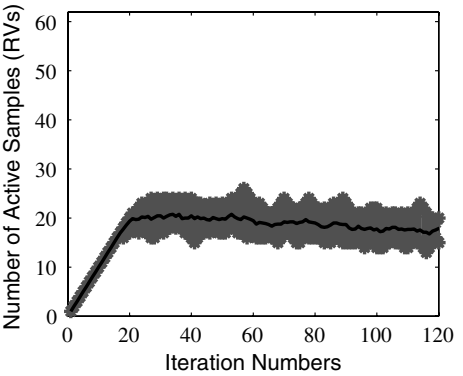


**Fig. 10** Number of retained RVs with $mRVM_1$.

iteration progresses. The number of RVs always keeps increasing during the first 20 iterations, and then the number of retained RVs seems to drop with a slight tendency. Meanwhile, the fluctuation in number of retained RVs is also observed. After 120 iterations, the minimal, the mean, and the maximal number of RVs are 15, 17.89, and 20, respectively, i.e., the fluctuation range in the number of retained RVs is six, due to a different train/test split.

The confusion matrix is commonly used as a good measure for the multiclass discrimination performance.[3,6,28] Table 1 gives the confusion matrix associated with all the types of activities. Specifically, the data located at the intersection of the *i*'th row and the *j*'th column represents the percentage of class *i* activity being recognized as class *j*. In other words, the main diagonal of the confusion matrix relates to the recognition accuracy of different activities while the remaining cells correspond to the percentages of misclassification.

As has been shown from Table 1, the averaged recognition accuracy has reached about 94.5%. However, there is a clear discrepancy among different activities, i.e., some activities, such as "bend" and "wave 1" both own recognition rates of 100% whereas for other activities, such as "run" and "skip" the related recognition rates are below 90% (85% and 87%, respectively). In addition, the misclassification rates are distributed in an unbalanced way. For instance, the confusion mainly occurs in three pairs: "run" and "walk," "skip and run," and "walk and run" (the corresponding misclassification rates are 9%, 9%, and 8%, respectively), which can be explained by their relative similarity. For the other pairs, the individual misclassification rates are less than 6%. Meanwhile, it is worth noting that the values of the obtained confusion matrix are non-systematic, which may reveal that the boundary of decision function is somewhat biased and still has the potential to be further optimized. More importantly, the above results suggest that even better recognition accuracy may be expected if we devise the feature model more intentionally, i.e., paying more attention to the prominent pairs with high misclassification rates (refer to the confusion matrix).

#### 5.2.2 *Results with the mRVM$_2$*

For a fair comparison between $mRVM_1$ and $mRVM_2$, here we use the same kernel function and bandwidth parameter as in the previous section.

Figure 11 shows the recognition results of $mRVM_2$. It can be seen that the mean recognition rate increases rapidly in the first five iterations, and after five iterations it reaches up to 96.37%. After 120 iterations, the worst, the mean, and the best recognition rates are 96.7%, 98.2%, and 100%, respectively, i.e., the fluctuation in recognition rates caused by variations in the training set is relatively small (only about 3.3%).

Figure 12 shows the number of retained relevance vectors (RVs) associated with given iteration numbers. Since the $mRVM_2$ is based on the top-down strategy, the RVs start with the whole training set and subsequently prune down the most irrelevant samples during each iteration. The number of active samples decreases sharply during the first 10 iterations and then drops more slowly. After 120 iterations, the minimal, the mean, and the maximal number of RVs are 11, 12.33, and 13, respectively, i.e. the size of retained RVs even approaches the number of classes (this paper involves a total of 10 classes of activities), and the fluctuation range in it is as small as three. Figures 11 and 12 show that as the iteration progresses, the size of retained RVs gradually reduces

**Table 1** Confusion matrix with $mRVM_1$.

| | Bend | Jack | Jump | P-jump | Run | Side | Skip | Walk | Wave 1 | Wave 2 |
|---|---|---|---|---|---|---|---|---|---|---|
| Bend | 100 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Jack | 0.00 | 100 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Jump | 0.00 | 0.00 | 93.3 | 4.50 | 0.00 | 0.00 | 2.20 | 0.00 | 0.00 | 0.00 |
| P-jump | 0.00 | 0.00 | 0.00 | 100 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Run | 0.00 | 0.00 | 0.00 | 0.00 | 85.0 | 2.00 | 4.00 | 9.00 | 0.00 | 0.00 |
| Side | 0.00 | 0.00 | 0.00 | 0.00 | 3.30 | 91.1 | 0.00 | 5.50 | 0.00 | 0.00 |
| Skip | 0.00 | 0.00 | 1.00 | 0.00 | 9.00 | 0.00 | 87.0 | 3.00 | 0.00 | 0.00 |
| Walk | 0.00 | 0.00 | 0.00 | 0.00 | 8.00 | 2.00 | 0.00 | 90.0 | 0.00 | 0.00 |
| Wave 1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 100 | 0.00 |
| Wave 2 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.10 | 98.9 |

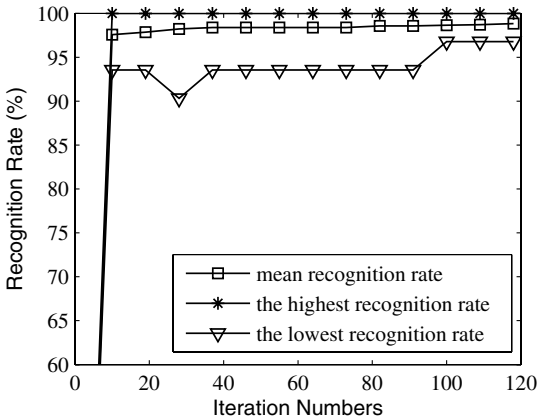


**Fig. 11** Recognition rate with $mRVM_2$.

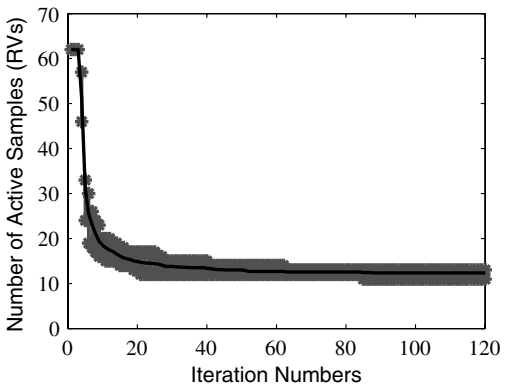


**Fig. 12** Number of retained RVs with $mRVM_2$.

while the mean recognition rate improves with each iteration. It demonstrates that the recognition rate deteriorates if the uninformative samples are included in the RVs.

Table 2 gives the resultant confusion matrix. The $mRVM_2$ achieves the average recognition rate of 98.2%, which is 3.5% higher compared with the $mRVM_1$. In addition, the most serious confusion occurs in two situations: the misclassification rates between "walk" and "run," "run and "walk" are 8% and 6%, respectively. These situations can be alleviated by introducing a velocity feature.

The experimental results show that the $mRVM_2$ outperforms the $mRVM_1$ both in terms of recognition rate and sparsity of RVs. However, further investigations are needed to examine the universality of this conclusion, i.e., experiments should be done on extensive human activity data sets.

### 5.3 *Comparative Evaluation*

In this section we first compare the recognition performances of our proposed feature representation and some other feature representation techniques. The evaluation is performed on the Weizmann data set, and the same classifier techniques—$mRVM_1$ and $mRVM_2$—are employed. The comparison results are listed in Table 3, which shows that the Zernike moments clearly relate to higher accuracy than the Hu moments. The reason may be that the Zernike moments have native rotational invariance and are far more robust to noise than Hu moments. It also shows the combination of Hu and Zernike can even achieve better recognition performance, which approximately equals the recognition accuracy associated with our proposed feature representation. However, our proposed features would be much simpler in implementation and thus more appropriate for real application when considering the relative high computational complexity accompanied with moment representations.

We subsequently compared the recognition performances between the mRVM classifiers and some other main classification techniques. The evaluation was also performed on the Weizmann data set, and the same feature model (VEI) was employed. In general, recognition accuracy is sensitive to the parameters of classifiers. For fair comparison, it is desirable to adopt optimal parameters for each kind of classifier. Specifically, as to the $k$-NN classifier, we set $k = 3$ since in this paper we can achieve the highest recognition accuracy under this condition. For the mRVM classifiers the Gaussian kernel was employed, and a nearly optimal bandwidth value (0.25) was adopted. For the SVM classifier, the Gaussian kernel was also employed, and the optimal

**Table 2** Confusion matrix with $mRVM_2$.

| | Bend | Jack | Jump | P-jump | Run | Side | Skip | Walk | Wave 1 | Wave 2 |
|---|---|---|---|---|---|---|---|---|---|---|
| Bend | 100 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Jack | 0.00 | 100 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Jump | 0.00 | 0.00 | 96.6 | 0.00 | 0.00 | 0.00 | 0.00 | 3.40 | 0.00 | 0.00 |
| P-jump | 0.00 | 0.00 | 0.00 | 100 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Run | 0.00 | 0.00 | 0.00 | 0.00 | 94.0 | 0.00 | 0.00 | 6.00 | 0.00 | 0.00 |
| Side | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 100 | 0.00 | 0.00 | 0.00 | 0.00 |
| Skip | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 100 | 0.00 | 0.00 | 0.00 |
| Walk | 0.00 | 0.00 | 0.00 | 0.00 | 8.00 | 0.00 | 0.00 | 92.0 | 0.00 | 0.00 |
| Wave 1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 100 | 0.00 |
| Wave 2 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 100 |

**Table 3** Comparison of recognition performance on Weizmann data set with the same classifier $mRVM_1$ or $mRVM_2$ but with different features.

| | Feature Representation | | | |
|---|---|---|---|---|
| Classifier | Hu | Zernike | Hu + Zernike | Our Features |
| $mRVM_1$ | 79.3 | 89.1 | 93.8 | 94.5 |
| $mRVM_2$ | 86.5 | 92.7 | 98.4 | 98.2 |

**Table 4** Comparison of recognition performance on Weizmann data set with the same features but different classifiers.

| | Different classifiers | | | | | |
|---|---|---|---|---|---|---|
| | NN | *k*-NN | Bayes | SVM | $mRVM_1$ | $mRVM_2$ |
| Accuracy (%) | 81.4 | 83.6 | 89.5 | 92.7 | 94.5 | 98.2 |

values of two controlling parameters, i.e., bandwidth and regularizing parameter $C$, were also fixed with the same searching strategy as in mRVM (in this paper, the optimal bandwidth and regularizing parameters are set to be 0.2 and 10, respectively). The comparison results are listed in Table 4, which shows that the mRVM classifiers, especially the $mRVM_2$, can yield much higher recognition accuracy than other classifier techniques.

Finally, we also compared the performance of our method with other existing methods on the same data set but with different classification techniques and different feature models. The quantitative results have been shown in Table 5. It can be seen that our method outperformed the other methods in terms of recognition rate, especially when the $mRVM_2$ was applied. Particularly, in the paper of Zhou et al.[29] both the SVM and the NN classifiers are employed (while the chosen feature models are absolutely the same), and the recognition performance of the former is much better than the latter. This consequence is as per our expectation since the SVM is a nonlinear classifier whereas NN is a linear classifier, i.e. the SVM inherently has stronger discriminative ability. Additionally, the pLSA classifier used in Ref. 30 is nonlinear. Our proposed feature extraction process is also simpler to implement. Specifically, the feature model in Ref. 29 includes 208 descriptive parameters that have been deduced from the time series data. In Ref. 30 the interest points are detected frame by frame to form a spatio-temporal cub, the dimension of which is subsequently reduced to 100 with the PCA algorithm. In Ref. 28 both the three-dimensional gradient and the orientation information are calculated at each pixel, and then a descriptor of 256 length and 2048 dimensions is generated. However, in our method the dimension of features is only 11, which is one or two orders smaller than the aforementioned methods. All the features are calculated merely from the spatio-temporal image (VEI or AME); thus, our method would be much less time-consuming during the feature extraction stage. In Ref. 13 the reported recognition accuracy value is approximate to that of $mRVM_1$ but clearly lower than that of $mRVM_2$. Moreover, that method needs to train multiple RVM classifiers for the multiclass problems. In contrast, the mRVM shows admirable sparsity and possesses the multiclass discriminative ability with a single classifier; thus, the computation cost during the classification stage is likely to be remarkably reduced. As a result, our method would have advantages in real-time applications.

In summary, the experiment results demonstrate that the mRVM exhibits considerable sparsity and strong discrimination in the multiclass activity recognition, which reveals its applicability to large-scale activity recognition problems. Furthermore, the proposed method is quite promising for real-time applications because: a. the mRVM classifiers, especially the $mRVM_2$, show considerable sparsity of RVs, i.e., the size of retained RVs even approaches the number of classes, which means the computation cost during the classification stage can be sharply reduced, b. the proposed features are quite simple to extract, and, meanwhile, only frames in a single motion cycle have to be dealt with, i.e., the feature extraction process also has relatively low computation complexity, and c. by using the mRVM, the classification procedure becomes more compact, efficient, and stable. Specifically, as for the traditional multiclass kernel machine learning algorithms, cascading of classifiers is always unavoidable. For instance, in Ref. 30 the multiple classifiers are composed of $C$ SVM classifiers based on the one-against-all strategy. In Ref. 29, the multiple classifiers are constructed

**Table 5** Comparison of recognition performance on Weizmann data set with different methods.

| Methods | Feature Extraction | Classifier | Accuracy (%) |
|---|---|---|---|
| our method | VEI model | $mRVM_1$ | 94.5 |
| | VEI model | $mRVM_2$ | 98.2 |
| Niebles et al.[30] | spatial-temporal interest points | pLSA | 90.0 |
| Scvanner et al.[28] | three-dimensional SIFT description | SVM | 84.2 |
| Zhou et al.[29] | structure-based statistical feature | NN | 78.3 |
| | structure-based statistical features | SVM | 91.4 |
| Yogameena et al.[13] | shape discriptions of silhouette points | multiple RVMs | 94.67 |

from $C(C-1)/2$ SVM classifiers based on the one-against-one strategy. By contrast, in the mRVM a single classifier is sufficient to discriminate all the classes.

## 6 Conclusion

In this paper we used mRVM classification technology to recognize human activities. In addition, a new spatio-temporal template was put forward, from which the activity features could be extracted in an efficient way. To validate the recognition performance of the proposed method, the 10 times three-fold cross-validation was carried out on the Weizmann data set, and the experiment results were given. Meanwhile, the recognition performance of other methods was also listed out for comparison. The results demonstrate that the mRVM classifiers, especially the $mRVM_2$, are ideal tools for activity recognition problems and have the potential to be applied on large-scale data sets.

There are mainly two aspects that need further improvements in future work. First, more flexibility can be added to the kernel function during the training stage. In the proposed method, the Gaussian kernel function with fixed bandwidth has been used for simplicity. The other type of kernel functions can also be considered. Meanwhile, the kernel parameter learning scheme[31] will be incorporated into the training stage so as to further improve the classification performance of our method. Second, we will put more emphasis on the design of the feature model. Misclassification situations still exist with our method, and the misclassification rates are distributed in an unbalanced way. Accordingly, if we devise the feature model more intentionally, the specific style of misclassification might be alleviated. For instance, the velocity feature might be powerful in distinguishing "run" from "walk." On the other hand, all the activities involved in this paper have strong periodicity. However, in practice some activities may be non-periodic at the scale of shot time, or they may be divided into a series of basic activities. Certainly, these situations would require variations in the feature model. For instance, we can modify the VEI by adding motion-direction information so as to strengthen the adaptability of our method to compound activity recognition. Additionally, extension of our work to the multi-view situation is also meaningful.

### Acknowledgments

The authors would like to thank Blank for providing the Weizmann data set and Psorakis et al. for sharing the theory of mRVM. Many thanks also go to the PhD candidates Raj Kumar Gupta and Chaoying Tang in Nanyang Technological University and to all the reviewers for their very helpful comments and suggestions. Our work has been supported by the Key Project of Chinese Ministry of Education (Grant No. 108174) and the Chongqing Municipal Natural Science Foundation of China (Grant No. CSTC2008BB3169).

### References

1. P. Antonakaki, D. Kosmopoulos, and S. J. Perantonis, "Detecting abnormal human behaviour using multiple cameras," *Signal Process.* **89**(9), 1723–1738 (2009).
2. W. L. Lu, K. Okuma, and J. J. Little, "Tracking and recognizing actions of multiple hockey players using the boosted particle filter," *Image Vis. Comput.* **27**(1–2), 189–205 (2009).
3. H. Meng, N. Pears, and C. Bailey, "A human action recognition system for embedded computer vision application," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1–6, IEEE Computer Society(2007).
4. B. Gholami, W. M. Haddad, and A. R. Tannenbaum, "Relevance vector machine learning for neonate pain intensity assessment using digital imaging," *IEEE Trans. Biomed. Eng.* **57**(6), 1457–1466 (2010).
5. N. Ikizler and P. Duygulu, "Histogram of oriented rectangles: a new pose descriptor for human action recognition," *Image Vis. Comput.* **27**(10), 1515–1526 (2009).
6. H. S. Chen et al., "Human action recognition using star skeleton," in *Proc. the 4th ACM International Multimedia Conference and Exhibition*, pp. 171–178, Association for Computing Machinery (2006).
7. 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).
8. L. Wang and D. Suter, "Informative shape representations for human action recognition," in *Proc. the 18th International Conference on Patten Recognition (ICPR 06)*, pp. 1266–1269, IEEE (2006).
9. D. Selvathi, R. S. Prakash Ram, and S. Selvi Thamarai, "Performance evaluation of kernel based techniques for brain MRI data classification," in *Proc. the International Conference on Computational Intelligence and Multimedia Applications (ICCIMA)*, pp. 456–460, IEEE Computer Society (2008).
10. A. M. Bogdanova, D. Nikolik, and L. Curfs, "Probabilistic SVM outputs for pattern recognition using analytical geometry," *Neurocomputing* **62**, 293–303 (2004).
11. J. Platt, "Probabilistic outputs for support vector machines and comparison to regularize likelihood methods," in *Advances in Large Margin Classifiers*, MIT Press, Cambridge, pp. 61–74 (2000).
12. M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.* **1**, 211–244 (2001).
13. B. Yogameena et al., "Human behavior classification using multi-class relevance vector machine," *J. Comput. Sci.* **6**(9), 1021–1026 (2010).
14. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag, NewYork (2006).
15. T. Damoulas et al., "Inferring sparse kernel combinations and relevance vectors: an application to subcellular localization of proteins," in *Proc. the 7th International Conference on Machine Learing and Applications*, pp. 577–582, IEEE Computer Society (2008).
16. I. Psorakis, T. Damoulas, and M. A. Girolami, "Multiclass relevance vector machines: sparsity and accuracy," *IEEE Trans. Neural Netw.* **21**(10), 1588–1598 (2010).
17. R. Poppe, "A survey on vision-based human action recognition," *Image Vis. Comput.* **28**(6), 976–990 (2010).
18. M. A. Ahad et al., "Moment-based human motion recognition from the representation of DMHI templates," in *Proc. the SICE Annual Conference-International Conference on Instrmentation, Control and Information Technology*, pp. 578–583, Society of Instrument and Control Engineers (2008).
19. C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in: *Proc. the 17th International Conference on Pattern Recognition*, pp. 32–36, IEEE (2004).
20. D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Comput. Vis. Image Understand.* **104**(2–3), 249–257 (2006).
21. H. Meng and N. Pears, "Descriptive temporal template features for visual motion recognition," *Pattern Recogn. Lett.* **30**(12), 1049–1058 (2009).
22. H. Qian et al., "Recognition of human activities using SVM multi-class classifier," *Pattern Recogn. Lett.* **31**(2), 100–111 (2010).
23. A. Elgammal, D. Harwood, and L. Davis, "Non-parametric model for background subtraction," in *Proc. the 6th European Conference on Computer Vision*, pp. 751–767, Springer Press (2000).
24. Y. Yacob and M. J. Black, "Parameterized modeling and recognition of activities," *Comput. Vis. Image Understand.* **73**(2), 232–247 (1999).
25. H. W. Lamn Toby, K. H. Cheung, and N. K. Liu James, "Gait flow image: a silhouette-based gait representation for human identification," *Pattern Recogn.* **44**(4), 973–987 (2011).
26. S. T. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science* **290**, 2323–2326 (2000).
27. L. Gorelick et al., "Actions as space-time shapes," *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(12), 2247–2253 (2007).
28. P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proc. the 5th ACM International conference on Multimedia*, pp. 357–360, Association for Computing Machinery (2007).
29. H. Zhou, L. Wang, and D. Suter, "Human action recognition by feature-reduced Gaussian process classification," *Pattern Recogn. Lett.* **30**(12), 1059–1066 (2009).
30. J. C. Niebles, H. Wang, and F. Li, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vis.* **79**(3), 299–318 (2008).
31. D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "Sparse Bayesian modeling with adaptive kernel learning," *IEEE Trans. Neural Netw.* **20**(6), 926–937 (2009).

**Weihua He** received aBSc degree in information engineering and anMSc degree in optoelectronic engineering from Chongqing University, China, in 2004 and 2007, respectively. She is currently working toward a PhD in Instrument Science and Technology at Chongqing University. Her current research interest is in machine learning and action recognition.

**Kin Choong Yow** received a BS degree from the National University of Singapore in 1993 and a PhD from Cambridge University in 1998. Currently, he is a professor at the Nanyang Technological University, Singapore. His research interest is in ambient intelligence, which includes passive remote sensing and computational intelligence.

**Yongcai Guo** received a PhD in instrument science and technology from Chongqing University, Chongqing, China, in 1999. Her research interests include signal processing and pattern recognition.

Photographs of the authors are not available.