Optical Engineering

OpticalEngineering.SPIEDigitalLibrary.org

Human action recognition using motion energy template

Yanhua Shao
Yongcai Guo
Chao Gao

SPIE.

Human action recognition using motion energy template

Yanhua Shao,[a,b,*] Yongcai Guo,[a] and Chao Gao[a]
[a]Chongqing University, Key Laboratory of Optoelectronic Technology and Systems of the Education Ministry, Chongqing 400030, China
[b]Southwest University of Science and Technology, School of Information and Engineering, Mianyang Sichuan 621010, China

Abstract. Human action recognition is an active and interesting research topic in computer vision and pattern recognition field that is widely used in the real world. We proposed an approach for human activity analysis based on motion energy template (MET), a new high-level representation of video. The main idea for the MET model is that human actions could be expressed as the composition of motion energy acquired in a three-dimensional (3-D) space-time volume by using a filter bank. The motion energies were directly computed from raw video sequences, thus some problems, such as object location and segmentation, etc., are definitely avoided. Another important competitive merit of this MET method is its insensitivity to gender, hair, and clothing. We extract MET features by using the Bhattacharyya coefficient to measure the motion energy similarity between the action template video and the tested video, and then the 3-D max-pooling. Using these features as input to the support vector machine, extensive experiments on two benchmark datasets, Weizmann and KTH, were carried out. Compared with other state-of-the-art approaches, such as variation energy image, dynamic templates and local motion pattern descriptors, the experimental results demonstrate that our MET model is competitive and promising. [DOI: 10.1117/1.OE.54.6.063107]

Keywords: human action recognition; action representation; filter; support vector machine; motion energy template.

Paper 141886 received Dec. 10, 2014; accepted for publication Jun. 3, 2015; published online Jun. 29, 2015.

1 Introduction

In recent years, automatic capture, analysis and recognition of human actions is a highly active and significant area in the computer vision research field, with plentiful applications both offline and online,[1,2] for instance, video indexing and browsing, automatic surveillance[3] in shopping malls, and smart homes, etc. Moreover, interactive applications, such as human-interactive games,[4] also benefit from the progress of human action recognition (HAR).

In this paper, we address the problem of representation and recognition of human activities directly from original image sequences. An action video can be interpreted as a three-dimensional (3-D) space-time volume (X-Y-T) by concatenating each two-dimensional (2-D) (X-Y) frame along one-dimensional time (T). Various literature demonstrates that spatiotemporal features, which include motion features and shape features, are elementary and useful for HAR.[2,5] This means that shape features and motion features are often combined to achieve more useful action representation.

The optical flow,[6] which is extracted from the motion between two adjacent image frames, can be utilized to distinguish action representation. Nevertheless, optical flow-based methods, such as histograms of optical flow[7] and motion flow history,[8] are affected by uncontrolled illumination conditions.

Another important class of action representation is based on gradients, such as histograms of oriented gradients (HOG)[9]. The HOG descriptor is capable of describing local edge structure or the appearance of an object, it is computed from the local distribution of gradients, and its performance is robust. However, gradient-based algorithms are sensitive to noise.

Many "shape features" of action in 3-D X-Y-T space are widely used in human action representation and HAR, for instance, the motion energy image (MEI)[6] and motion history image (MHI).[10] However, those methods are not immune to motion cycles.

Based on the idea that an action can be considered as a conglomeration of motion energy in a 3-D space-time volume (X-Y-T), which is treated as an "action-space," we introduce a new high-level semantically rich representation model, which is called motion energy template (MET) model, that is based on the filter bank for HAR. It should be stressed that similar filter-based methods have been applied with success to other challenging video understanding tasks, e.g., spacetime stereo,[11] motion estimation,[12,13] and dynamic scene understanding analysis.[14] The framework of our method is shown in Fig. 1. The MET method, which is illuminated by the object bank method[15] and action spotting,[16] performs recognition by template matching. The MET model is obtained directly from video data, so some limitations of classical methods can be avoided, such as foreground/background segmentation, prior learning of actions, motion estimation, human localization and tracking, etc. Taking the silhouette-based method for example, background estimation is an important and challenging task to improve the quality of silhouette extraction.[17]

Input videos typically consist of template videos and search videos (unrecognized candidate videos), as shown in Fig. 1. In our method, the motion template is defined first by a small template video clip. Human actions are expressed as the composition of motion energy in a high-dimensional "action-space" in several predetermined spatiotemporal orientations by 3-D filter sets. In other words, the representation task is achieved by the MET model, and the classification task is fulfilled directly by a classifier [such as support vector machine (SVM)]. The algorithm processes, as shown in

*Address all correspondence to: Yanhua Shao, E-mail: syh@cqu.edu.cn

Fig. 1, are as follows. (1) The 3-D Gaussian filter bank is used to decompose input videos into shorthand for space-time oriented motion energy (SOME) volumes (Sec. 3.1). (2) The SOME volumes are then matched to a database of the SOME template volumes at the corresponding spatio-temporal points using the Bhattacharyya coefficient. By this means, the similarity volumes of the action template (**T**) and the unrecognized video (**S**) are obtained (Sec. 3.2). (3) After 3-D max-pooling (3DMP), we get the MET features (Sec. 3.3). (4) Finally, the MET features can be used to obtain the action labels. In our experiments, by combining with a linear SVM on the benchmark datasets, i.e., Weizmann and KTH, our method achieves 100% and 95.37% promising accuracies, respectively.

Our contributions could be summarized as follows:

(1) We proposed a novel template-based MET algorithm which could generate discriminative features directly from the video data for HAR.

(2) We evaluated the MET model on two benchmark action datasets and showed that MET model is an appreciative tool for action representation, enabling us to obtain the highest reported results on the Weizmann dataset.

(3) We demonstrated that our method achieves excellent results on a benchmark dataset (KTH) despite the different scenarios and clothes.

The remainder of this paper is organized as follows. In Sec. 2, we briefly review the related work in the field of HAR. In Sec. 3, we elaborate on the MET model. In Sec. 4, we present the experimental results from two public benchmark action recognition datasets, Weizmann and KTH. Finally, conclusions are given in Sec. 5.

2 Related Work

HAR is often done in two steps: action representation and action classification. The first key step is action representation. There exists a great deal of literature on human action representation and HAR.[5,18] In this section, we focus discussions mainly on action representation, especially high-level features and template-based methods, which are more relevant to our approach.

2.1 *High-Level Features*

In spite of a robust low-level image, features have been proven to be effective for many different kinds of visual recognition tasks. However, for some high-level visual tasks such as scene classification and HAR, many low-level image representations carrying relatively little semantic meaning are potentially not good enough. Object bank[15] was proposed as a new "high-level image representation" based on filter banks for image scene classification. Action spotting,[16] a novel compact high-level semantically rich representation method, was introduced based on the space-time oriented structure representation. Those methods carry relatively more semantic meaning.

2.2 *Template-Based Method*

The "template-based" method has gained increasing interest because of its convenience for the computing process.[6,19–21]

Bobick and Davis[6] computed Hu moments of MEI and MHI to create action templates based on a set of training examples. Kellokumpu et al.[22] proposed a new method using texture-based feature work with raw image data rather than silhouettes. Dou and Li[20] constructed motion temporal templates by combining the 3-D scale invariant feature transform with the templates of Bobick.

Chaudhry et al.[19] modeled the temporal evolution of the object's appearance/motion using a linear dynamical system from sample videos and used the models as a dynamic template for tracking objects in novel videos.

Efros et al.[23] proposed a template-based method based on optical flow. Their methods can be thought of as a special type of action database query and are effective for video retrieval tasks.

Shechtman and Irani[24] used a behavior-based similarity measure to extend the notion of the traditional 2-D image correlation into 3-D space-time video-template correlation and they further proved that the 3-D correlation method has good robustness to small changes in scale and orientation of the correlated behavior.

Hae Jong and Milanfar[21] introduced a novel method based on the matrix cosine similarity measure for action recognition. They used a query template to find similar matches.

It should be noted that despite the fact that these methods are all based on the template pattern, the action representations obtained are relatively varied. For instance, some methods may require background estimation,[6,25] noise reduction, period estimation,[25] object segmentation, human localization or tracking,[23] and so on. These pretreatments may not be conducive to automatically recognize action in real applications.

2.3 *Action Classification*

Classifier is an important factor which affects the performance of HAR. Heretofore, many famous pattern classification techniques [for instance, k-nearest neighbor,[10] probabilistic latent semantic analysis (pLSA),[26] neural network (NN,[17] SVM,[21,27] relevance vector machine (RVM),[25,28] and multiple kernel learning][29] and their modifications have been proposed and employed in the action recognition field.

More detailed surveys on action recognition can be found in Refs. 2 and 18.

3 MET Model

As mentioned before, Fig. 1 shows the framework of our method, which consists of the following three algorithmic modules: (1) filtering, (2) measuring SOME volumes similarity based on the Bhattacharyya coefficient, and (3) 3DMP. Finally, after achieving the above steps, the MET features were gained. Later in this section, we will elaborate on each step from Secs. 3.1–3.3, respectively.

3.1 *SOME Features Construction for MET Model: Filtering*

The requested space-time oriented decomposition is obtained by the phase-sensitive third derivative of 3-D Gaussian filters[13,30,31] $G_{3_{\hat{\theta}}}(\mathbf{x}) \equiv \partial^3 k \exp[-(x^2+y^2+t^2)]/\partial\hat{\theta}^3$, with $\mathbf{x} = (x, y, t)$ denoting the space-time position and k is a normalization factor. $\hat{\theta} \equiv (\alpha, \beta, \gamma)$ is the unit vector capturing

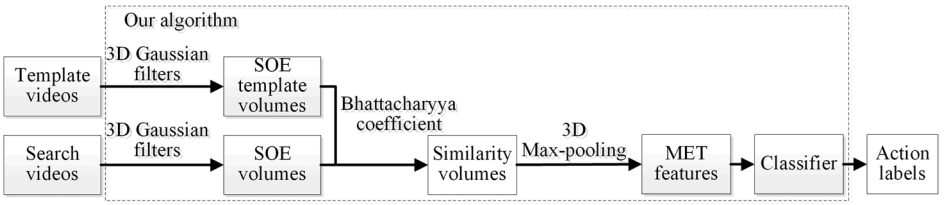

Fig. 1 Framework of our action recognition system which consists of the following three algorithmic modules: filtering, measuring shorthand for space-time oriented motion energy volumes similarity based on Bhattacharyya coefficient and three-dimensional max-pooling (3DMP).

their 3-D direction of the filter symmetry axis and α, β, γ are the direction cosines according to which the orientation of the 3-D filter kernel is steered.[31] More detailed expositions on the mathematical formulation and design of the filters can be found in Refs. 30 and 31. (The filter code can be obtained by email for academic research).

A locally summed pointwise energy measurement can be gained by rectifying the responses of the raw video to those filters over a visual space-time neighborhood $\Omega(\mathbf{x})$, which covers the entire action of the video sample under analysis, as follows:

$$E_{\hat{\theta}}(\mathbf{x}) = \sum_{\mathbf{x}\in\Omega(\mathbf{x})} (G_{3_{\hat{\theta}}} * I_{\text{in}})^2, \tag{1}$$

where $*$ denotes convolution and I_{in} is the input video. Spatiotemporally oriented filters are phase sensitive,[13] which is to say that the filters' output may be positive, negative, or zero, so that the instantaneous output does not directly signal the motion.[12] However, by squaring and summing those filters' outputs, this process follows from Parseval's theorem and the resulting signal gives a phase-independent measure of motion energy which is always positive and directly signals the motion.[12]

$$E_{\hat{\theta}}(\mathbf{x}) \propto \sum_{\omega_x,\omega_y,\omega_t} |\mathcal{F}\{G_{3_{\hat{\theta}}} * I_{\text{in}}\}(\omega_x, \omega_y, \omega_t)|^2, \tag{2}$$

where $\mathcal{F}$ denotes the Fourier transform, (ω_x, ω_y) is the spatial frequency and ω_t signifies the temporal frequency.

Obviously, retaining both the visual-spatial information and dynamic behavior of the action process in the region of interest, which is determined by filtering, an example of which is illustrated in Fig. 2(b), is relatively important. But it is unnecessary and redundant to describe the detailed differences among different people, who perform the same action wearing different clothes in several different scenarios. That is, in action recognition, the dynamic properties of an actor are more important than the spatial appearance, which comes from the actors' different clothes, etc. Nevertheless, in the MET method, human actions are expressed as the composition of motion energy along with several predetermined spatiotemporal orientations, θ, and the responses of the ensemble of oriented energy are partly appearance dependent. To overcome this issue, we take full advantage of 3-D marginal information, which emphasizes and highlights the value of the dynamic properties in the process of building the spatial orientation component. The process can be reformulated with more detail as follows.

As is well known, when using an N-order 3-D derivative of Gaussian filters, $(N+1)$ directional channels are required to span in a reference plane.[30] In this study, $N = 3$ is adopted in the process of space-time oriented filtering which is defined as in Eq. (1). Consequently, it is suitable to consider four directional channels along each reference plane in the Fourier domain. Finally, we obtained a group of four isometric directions within the plane:

$$\begin{aligned}
\hat{\theta}_i &= \cos\left(\frac{\pi i}{4}\right)\hat{\theta}_a(\hat{n}) + \sin\left(\frac{\pi i}{4}\right)\hat{\theta}_b(\hat{n}),\\
\hat{\theta}_a(\hat{n}) &= \hat{n} \times \hat{e}_x / \|\hat{n} \times \hat{e}_x\|,\\
\hat{\theta}_b(\hat{n}) &= \hat{n} \times \hat{\theta}_a(\hat{n}),
\end{aligned} \tag{3}$$

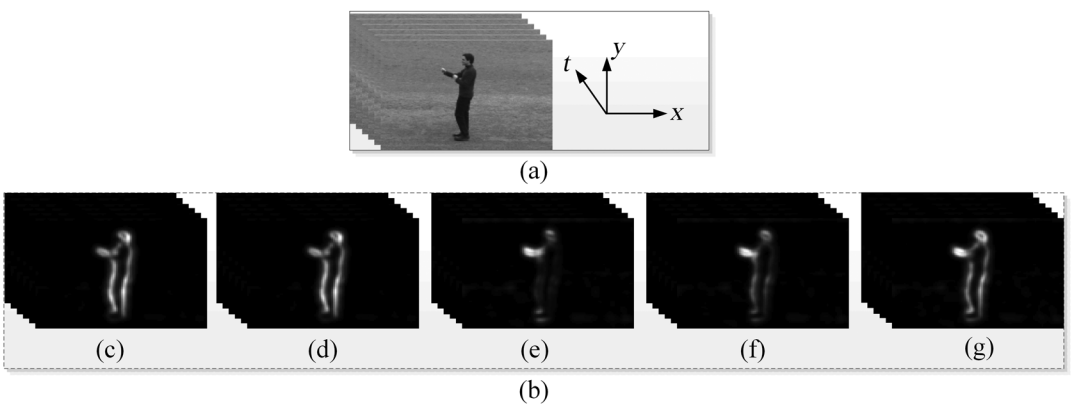

Fig. 2 General structure of the motion energy representation. (a) Input video ($x \times y \times t = 160 \times 120 \times 360$): boxing taken from the KTH action dataset. (b) Oriented motion energy volumes. Five different space-time orientations are made explicitly. (c) downward motion; (d) upward motion; (e) leftward motion; (f) rightward motion; (g) flicker motion.

where $0 \le i \le 3$, $\hat{n}$ signifies the unit normal of a frequency domain plane and $\hat{e}_x$ is the unit vector along the ω_x axis.

Now, finally, the marginalized motion energy measurement along a Fourier domain plane can be obtained by summing the energy measurements, $E_{\hat{\theta}_i}$, in all four specified predefined directions. Those directions are typically expressed as $\hat{\theta}_i$ in Eq. (3).

$$\tilde{E}_{\hat{n}}(\mathbf{x}) = \sum_{i=0}^{3} E_{\hat{\theta}_i}(\mathbf{x}). \tag{4}$$

Each $E_{\hat{\theta}_i}$, as shown in Eq. (4), is computed by Eq. (1). In the present implementation, five energy measurements of an action are made explicitly at several different directions, $\hat{\theta}$. Finally, the normalized energy measurement is composed of the energy of each channel responses at each pixel by

$$\hat{E}_{\hat{n}_i}(\mathbf{x}) = \tilde{E}_{\hat{n}_i}(\mathbf{x}) / (\sum_{j=1}^{5} \tilde{E}_{\hat{n}_j}(\mathbf{x}) + \varepsilon), \tag{5}$$

where ε, which depends on the particular action scenario, is a constant background noise for avoiding instabilities at the space-time position where the entire motion energy is too small. By using Eqs. (3) and (4), we obtained five normalized SOME measurements.

Based on the above-mentioned theories, for clarity, we present a pictorial display of the general structure of the space-time oriented structure representation for the MET model, as shown in Fig. 2. Figure 2(a) shows an example of a 3-D X-Y-T volume corresponding to the human action of boxing. Each oriented motion energy measurement is extracted from the response to the oriented motion energy filtering along a predefined spatiotemporal orientation, $\hat{\theta}$, as shown in Fig. 2(b), corresponding to leftward $(-1/\sqrt{2}, 0, 1/\sqrt{2})$, rightward $(1/\sqrt{2}, 0, 1/\sqrt{2})$, upward $(0, 1/\sqrt{2}, 1/\sqrt{2})$, downward $(0, -1/\sqrt{2}, 1/\sqrt{2})$, and flicker $(1, 0, 0)$ motion.

3.2 *Measuring SOME Volumes Similarity: Template Matching*

After obtaining the SOME template volumes and SOME volumes of the search videos, similarity calculation is required in order to get the MET features.

In order to define a (dis)similarity measure between probability distributions, a variety of information-theoretic measures can be used.[14,32,33] It was demonstrated that in numerous practical applications, the Bhattacharyya coefficient provided better results as compared to the other related measures (such as Kullback–Leibler divergence, L_1, L_2, etc.).[14,32] Furthermore, there is also a "technical" advantage gained from using the Bhattacharyya coefficient, which has a particularly simple analytical form.[32]

Therefore, here, we use the Bhattacharyya coefficient $m(\cdot)$, which is robust to small outliers, for motion energy volumes similarity measurement. The range of this measure is [0, 1]. Herein, 0 indicates complete disagreement, in-between values indicate higher similarity, and 1 denotes absolute agreement. The individual histogram similarity measurements[33] are expressed as a set of Bhattacharyya coefficients.

As mentioned above, the MET $\mathbf{T}$ is usually defined by small template action video clips and $\mathbf{S}$ signifies the search video. The global match measurement, $M(\mathbf{x})$, is represented by

$$M(\mathbf{x}) = \sum_{\mathbf{r}} m[\mathbf{S}(\mathbf{r}), \mathbf{T}(\mathbf{r} - \mathbf{x})], \tag{6}$$

where $\mathbf{r} = (u, v, w)$ denotes the range of the predefined template volume. Hence, $m[S(\mathbf{r}), T(\mathbf{r} - \mathbf{x})]$, which signifies the similarity between $\mathbf{T}$ and $\mathbf{S}$ at each space-time position, is summed across the predefined template volume. The global peaks of similarity measure roughly estimate the potential match locations.

In short, we obtained the similarity volumes by using a Bhattacharyya coefficient-based template matching algorithm.

3.3 *MET Features' Vector Construction: 3DMP*

The similarity volumes were then used to calculate the MET features' vector through the 3DMP (MP) method. In specific, the 3DMP method[34,35] is used to calculate a similarity measurement with three levels in the octree (as shown in Fig. 3). We note that the 3DMP method has two remarkable properties for feature expression in the MET model: (1) 3DMP is able to generate a fixed-length output vector regardless of the input size of the similarity matrix/volume and (2) 3DMP uses multilevel spatial bins. Multilevel pooling has been shown to be robust to object deformations.[35,36] Therefore, this constructs a 73-dimension feature vector, $X = \{x_1, \ldots, x_{73}\}$, for each action pair.

In our implementation, 102 action videos were selected from the Weizmann[37] dataset and the KTH[27] dataset in order to construct the template set, in particular, 72 (three actors performed six actions under four scenarios) video clips from KTH and 30 (three actors performed 10 actions) video clips from Weizmann.

As for the scale problem, a scale has two aspects: spatial scale and temporal scale (motion periodicity). In particular, the spatial scale determines the size of the objects/actors to the most degree. In our implementation, the scale of the predefined template volume for the MET is unfixed. On the contrary, in order to improve the robustness of the MET model, we should take different spatial scales and temporal scales into consideration in the process of selecting a template. The influence on recognition results by different numbers of scales was analysed in Sec. 4.6 by verifying different numbers of MET.

In other words, $N_t = 102$. For the given MET model with N_t templates, we achieve N_t correlation volumes from the MET model. Hence, the overall length of the MET features' vector would be 7446 ($N_t \times 73 = 102 \times 73$). Thus, we have the MET features defined for classifiers input.

4 Experimental Results and Analysis

In this section, our approach is evaluated on two action recognition datasets, Weizmann[37] and KTH,[27] which are widely used as benchmarks. Our experiment was based on MATLAB code implemented on a 2.4 GHz Intel processor without special hardware acceleration (such as parallel computing, multicore CPUs, GPUs, etc.). The LIBSVM software is used to classify the actions.[38] A linear SVM classifier combined with the MET model defines one novel method for HAR. Sections 4.1 and 4.2 contain the comparative evaluation using Weizmann[37] and KTH,[27] respectively.

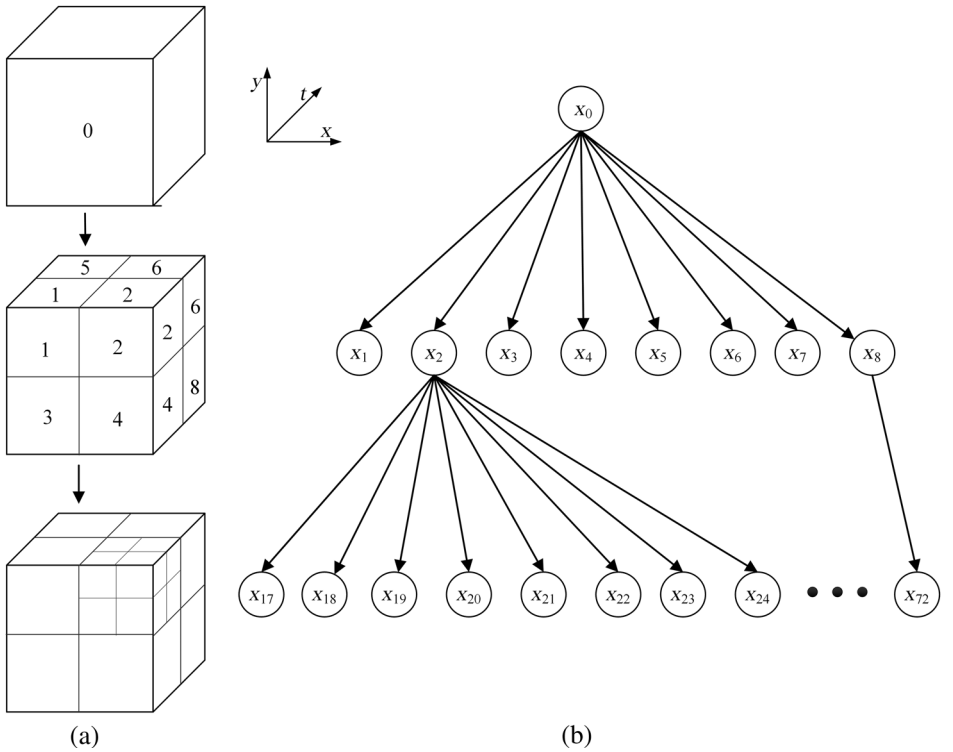

Fig. 3 The schematic of 3DMP. (a) Recursive subdivision of a cube into octants. (b) The corresponding octree.

Abundant information of the evaluation of the MET model is also given in Sec. 4.3–4.6. Specifically, the run time of the MET model is analyzed in Sec. 4.3. After that, we individually evaluate the impact of a classifier in Sec. 4.4. Then the performance with different dimensionality reduction methods is evaluated in Sec. 4.5. Finally, the performance with different numbers of METs is directly compared directly in Sec. 4.6.

4.1 *Action Recognition on Weizmann Dataset*

In this section, the proposed method is tested with a standard Weizmann benchmark dataset,[37] which provides a good platform for comparing the MET model with other methods under similar evaluation setups. Here, first, we will provide a brief introduction about this dataset. Then the experimental evaluation results and discussions are reported in Sec. 4.1.2.

4.1.1 *Weizmann dataset*

This single-scenario dataset involves 90 uncompressed colorful videos with an image frame size of 180×144 pixels $(25\ \text{frames}/\text{s})$. Figure 4 shows some example frames from this dataset.

4.1.2 *Experimental results and discussion*

In this section, 10 rounds of threefold cross validation are performed. The quantitative recognition performance and some state-of-the-art methods are shown in Table 1, such

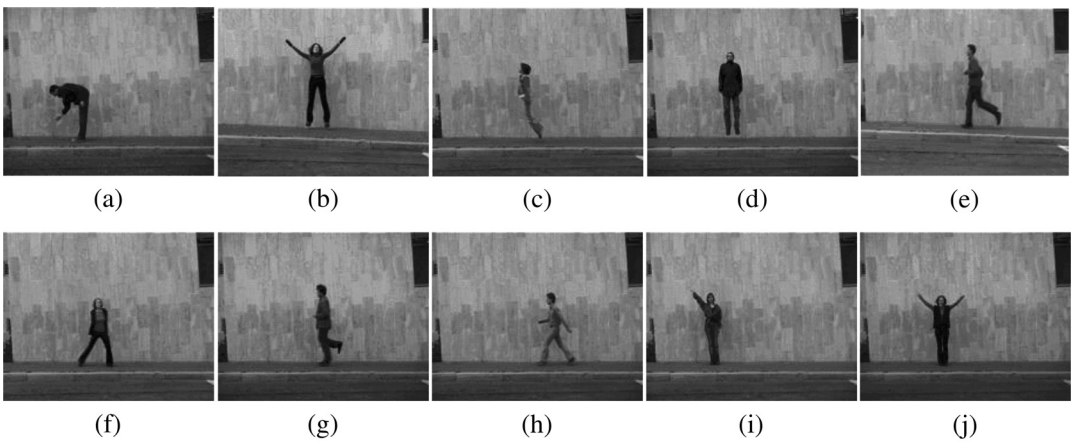

Fig. 4 Sample frames from the ten actions in the Weizmann dataset: (a) bend; (b) jack; (c) jump-forward; (d) jump-up-down; (e) run; (f) gallop-sideways; (g) skip; (h) walk; (i) wave-one-hand; (j) wave-two-hands.

Table 1 Comparing the recognition performance on the Weizmann dataset.

Approach	Feature expression	Classifier	Accuracy (%)
Niebles et al.[26]	Space-time interest points	pLSA	90
Zhou et al.[17]	Silhouettes	SVM	91.4
Chaudhry et al.[19]	Dynamic templates	*k*-NN	92.5
Yogameena et al.[28]	Shape descriptions of silhouette	RVM	94.6
Bouziane et al.[40]	3-D Zernike moments	Spectral graph	96.3
He et al.[25]	Variation energy image	mRVM	98.2
Guha et al.[39]	Local motion pattern descriptors	Concatenated dictionary	98.9
Our approach	MET model	SVM	100

Note: pLSA, probabilistic latent semantic analysis; SVM, support vector machine; k-NN, k-nearest neighbor; RVM, relevance vector machine; MET, motion energy template.

as the variation energy image (VEI) model,[25] dynamic templates[19] and local motion pattern descriptors.[39] Table 1 shows that RVM has a higher recognition accuracy than SVM when based on a similar feature expression.[17,28] The moment-based method provides a very useful analysis tool for HAR and obtains some satisfying results.[25,40] Rubner[33] explores the effectiveness of sparse representations for action recognition in videos. In Ref. 21, the VEI model is much less time consuming during the feature extraction stage. Nevertheless, this method requires silhouette extraction, period estimation and background. References 15 and 21 are based on the template pattern, but different feature extractions and classifiers are specifically employed.

Intuitively, the method we proposed has a higher recognition rate than these state-of-the-art methods. This is mainly due to the following three reasons: (1) the MET model is more effective; (2) the Weizmann dataset is not challenging enough because of its single static scenario; and (3) SVM, which is based on statistical principle, is one of the most successful classification techniques.

4.2 *Action Recognition on KTH Dataset*

4.2.1 *KTH dataset*

The KTH dataset[27] contains six human actions and each action is performed under four different scenarios which are not presented in action dataset Weizmann.[37] For this reason, it is a more challenging dataset. Figure 5 shows some sample frames of this dataset.

4.2.2 *Experimental results and discussion*

In this section, evaluation with the experimental setup is reported: the training set (eight subjects) and the test set (nine subjects). Here, we compared the performance of the

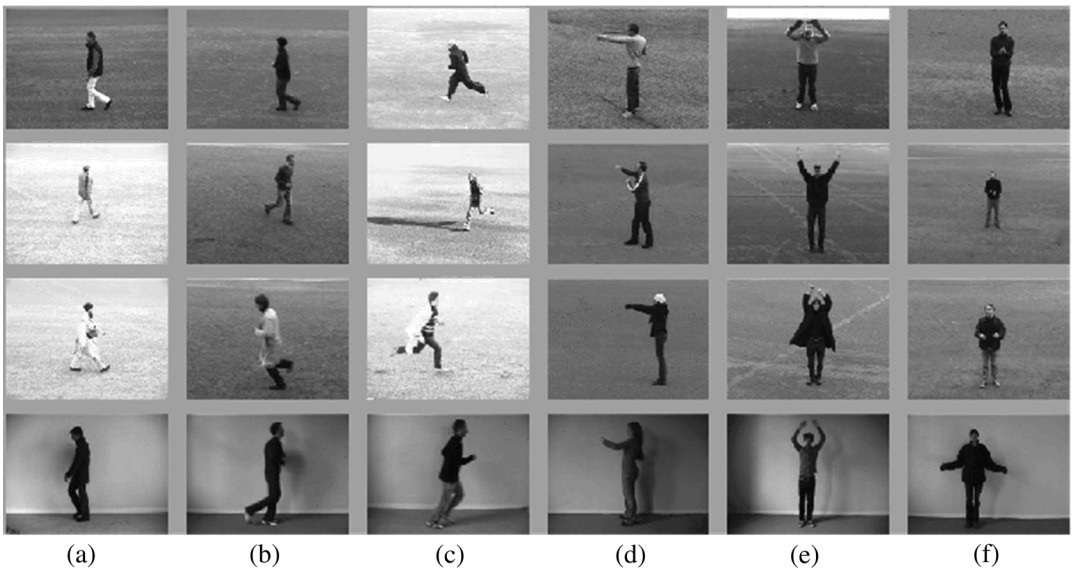

Fig. 5 Sample frames from the KTH dataset.[23] All six classes [columns, (a–f): walking, jogging, running, boxing, waving, and clapping] and four scenarios [rows, top to bottom: S1—outdoors, S2—outdoors with scale variation, S3—outdoors with different clothes, and S4—indoors] are presented.

Table 2 Recognition accuracies on the KTH dataset.

Algorithm	Feature extraction	Classifier	Accuracy (%)
Schuldt et al. [27]	Local space-time features	SVM	71.7%
Niebles et al. [26]	Space-time interest points	pLSA	81.5%
Derpanis et al.[16]	Spatiotemporal orientation analysis	*k*-NN	89.34%
Huang et al.[41]	Space-time interest points + optical flow	hCRF	89.7%
Kovashka et al.[29]	Hierarchical space-time feature	MKL	94.53%
Xinxiao et al.[42]	Spatiotemporal context and appearance	MKL	94.5%
Our method	MET model	SVM	95.37%

Note: pLSA, probabilistic latent semantic analysis; SVM, support vector machine; k-NN, k-nearest neighbour; hCRF, hidden conditional random field; MKL, multiple kernel learning; MET, motion energy template.

proposed method with other methods on the same benchmark dataset. The quantitative recognition results and some state-of-the-art methods are listed in Table 2. This shows that the proposed method has a significantly better performance in terms of average accuracy compared with the state-of-the-art methods. It should be noted that in many studies, only one dataset (Weizmann or KTH) is validated. Some studies (including our work) have shown that the recognition ratio on Weizmann dataset is higher than on the KTH based on the same methods.[21,26] More specifically, the method [space-time interest points (STIP) + pLSA], on the benchmark datasets, Weizmann and KTH, achieves 90% and 71.7% accuracies, respectively. In our experiments on Weizmann and KTH, our method achieves 100% and 95.37% accuracies, respectively.

It is also interesting to note that in Ref. 41, a hidden conditional random field is more effective than SVM and HMM based on the same fusion feature (STIP + optical flow).

The confusion matrix is another commonly used evaluation method of the classification performance. The confusion matrix for our method is shown in Fig. 6. It is interesting to note that the major confusion occurs between the "hand clapping" and the "hand waving." This is partly due to the fact that both of them have close local motion appearance.

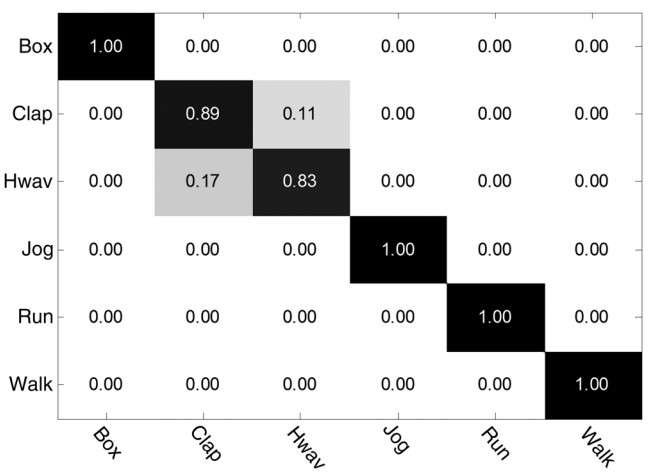

Fig. 6 Confusion matrix for KTH dataset.

It is clearly seen from Fig. 6 that "box," "jog," "run," and "walk" obtain a recognition rate of 100%. Hence, our approach can achieve a better performance by paying more attention to the misclassified activities as mentioned above.

4.3 *Run Time of MET Method*

In many real applications, computation cost is one of the critical factors. Here, we give a quantitative analysis for computation cost. From the viewpoint of mathematics, measuring motion energy similarity is convolution.[12] The MATLAB program runs on an Intel 2.4 GHz machine without special hardware acceleration (such as GPU). A video clip (name: daria_jump.avi, columns: 180, rows: 144, frames: 67) from the Weizmann dataset is taken as an example, and the total elapsed time is 2611.2 s ($N_t = 102$, i.e., 25.6 s is the average time for each template). Thus, the MET method takes 4.2639 s for calculating the motion energies stage. This is essentially due to the low computational complexity of 3DMP.[35,36] Much of the time required to build a new MET feature is spent on template matching (measuring SOME volumes' similarity). Also note that our method returns not only the similarity, but also the locations in the video where the query clip is matched if needed.

In our implementation, the overall search strategy is adopted. There are some other strategies to deal with template matching, such as coarser sampling and coarse-to-fine strategy.[43] Ning et al.[43] introduced a coarse-to-fine search and verification scheme for matching. In their coarse-to-fine strategy, the searching process takes about one-ninth of the time to scan the entire video. However, coarse-to-fine search algorithms have some probability of misdetection.[24]

Above all, however, measuring motion energy similarity could be easily implemented using multithreading and parallel-processing techniques for minimizing the "time to consume," because most of the computation involves convolution. Here, an example included in the compute unified device architecture (CUDA) SDK was used to illustrate the quantitative analysis. The CUDA conv performed 2-D convolution using an NVIDIA graphics chipset that can outperform conv2, which was a built-in function from MATLAB, by as much as 5000%. It is expected that parallel processing will significantly improve the speed in real applications.

Table 3 Comparing the recognition performance on the KTH dataset with different classifiers.

	Different classifier			
	BP-NN	Bayes	*k*-NN	SVM
Accuracy (%)	77.31	91.67	93.98	95.37

Note: SVM, support vector machine; k-NN, k-nearest neighbor.

4.4 *Varying the Classifiers*

We compared the recognition performances on the KTH dataset between the SVM classifier and some other mainstream classification methods and the same MET model was employed. To compare all methods fairly, the optimal parameters, which have been optimized by cross validation of these classifiers were employed. For instance, in the case of the backpropagation NN classifier, the number of hidden layer nodes is set to 10. 3-NN is adopted. For the SVM classifier, the linear kernel was adopted. The comparison results were shown in Table 3. Intuitively, the SVM classifier can acquire a higher recognition rate than that of other classifiers.

4.5 *Varying Dimensionality Reduction Techniques*

For high-dimensional dataset/features (i.e., in this paper, our MET model with number of dimensions $d_{\text{all}} = 7446$), dimension reduction is usually performed prior to applying a classifier in order to (1) prevent the problems derived from the curse of dimensionality and (2) reduce the computing time in the stage of classification.

In recent years, a variety of dimensionality reduction techniques have been proposed for solving this problem, such as principal component analysis (PCA), kernel PCA (KPCA), the linear discriminant analysis (LDA), generalized discriminant analysis (GDA),[44] locally linear embedding (LLE), and so on.

In general, dimensionality reduction techniques can be divided in two major categories: linear dimensionality reduction, such as PCA and LDA; and nonlinear dimensionality reduction, such as KPCA, GDA, and LLE. More detailed surveys on dimensionality reduction techniques can be found in Refs. 45 and 46.

Table 4 Comparing the recognition performance on the KTH dataset with different feature reduction techniques.

	Different classifier			
Reducer	3-NN	5-NN	Bayes	SVM
PCA	**94.91**	93.52	93.51	94.44
KPCA	**92.13**	90.74	87.96	89
LDA	100	100	100	100
GDA	100	100	99.53	100
LLE	91.20	90.74	**91.67**	89.35

Abbreviations: SVM, support vector machine; PCA, principal component analysis; KPCA, kernel principal component analysis; LDA, linear discriminant analysis; GDA, generalized discriminant analysis; LLE, locally linear embedding.
Note: The bold values represent the three best results for each method.

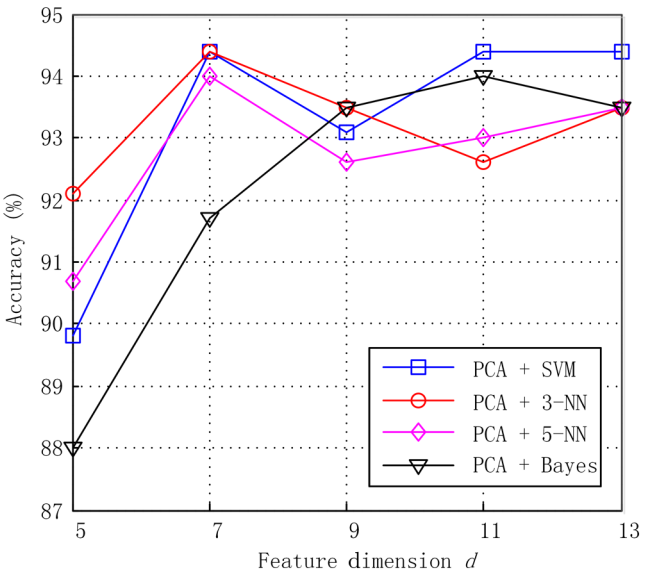

Fig. 7 Evaluation of the size of features on the KTH dataset.

The evaluation results are shown in Table 4, which shows that LDA and GDA can acquire a higher recognition rate than that of other reduction techniques. However, LDA and GDA are supervised subspace learning methods which use labels to choose projection. In real applications, we often face unlabeled data.

Moreover, the comparison of different kernel functions for KPCA shows that the linear kernel function is most suitable for our method.

Results for various specific feature dimensions, which are gained through PCA, are shown in Fig. 7. For all classification techniques, the performances are improved while the dimension d increases up to 7. The reason for this phenomenon is that the feature needs enough dimensions to encode action information.

4.6 *Varying the Number of METs*

From a mathematical perspective, the size of the MET model plays a crucial part in recognition performance and

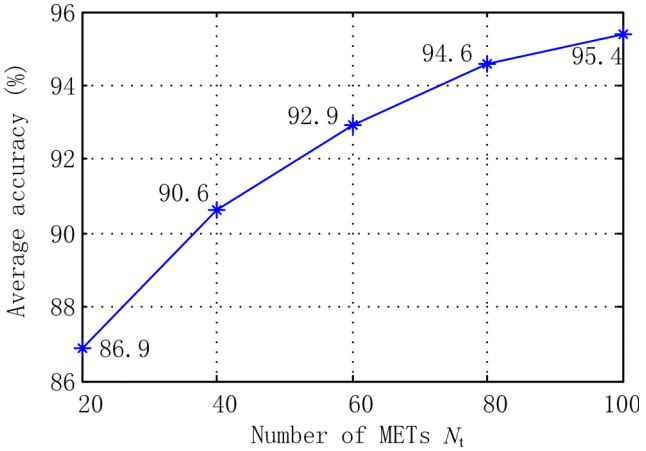

Fig. 8 Evaluation of the size of the motion energy template on the KTH dataset.

computational cost. In our approach, analyzing and evaluating the number of METs were also performed on the KTH dataset. For each different number N_t, we ran 100 iterations. It is arranged as follows: (1) we randomly select N_t motion template from all 102 METs and construct a new feature and (2) the evaluation is reported with the training set (eight subjects) and the test set (nine subjects).[27] The results are reported in Fig. 8. It can be seen that the recognition accuracy increases with a larger MET model. Improving the expression ability of the MET model depends on having sufficient templates, however, with more templates the total computing time will also increase. As previously mentioned, proper use of special hardware (such as parallel computing, multicore CPUs and GPUs, etc.) can remarkably accelerate computations for time-sensitive applications.

Contrasted with other methods, competitive experimental results are obtained with a relatively small number of METs. For example, with $N_t = 60$ on the benchmark dataset (KTH), our method achieves a 92.9% promising accuracy.

5 Conclusions

In this paper, a novel approach based on filter banks is presented for human action analysis by describing human actions with the MET model, a new high-level representation of video based on visual space-time oriented motion energy measurements. The MET model is achieved with the filter bank. In other words, actions are expressed as the composition of energy along with several predetermined spatiotemporal orientations in a high-dimensional "action-space" by filter sets. As the MET model is derived from raw image sequences data, many disadvantages, such as object location and segmentation, can be ignored. Moreover, the MET method is much less sensitive to spatial appearances such as hair and clothing. Extensive experiments on the Weizmann dataset and the KTH dataset have demonstrated that the MET model is an ideal method for the HAR problem and other video understanding tasks.

Acknowledgments

The authors would like to thank Schuldt et al. for providing the KTH dataset. The study was partly supported by the Key Project of Chinese Ministry of Education (grant No. 108174) and the PhD Programs Foundation of Ministry of Education of China (grant No. 20130191110021).

References

1. T. B. Moeslund, A. Hilton, and V. Kruger, "A survey of advances in vision-based human motion capture and analysis," *Comput. Vision Image Underst.* **104**(2–3), 90–126 (2006).
2. R. Poppe, "A survey on vision-based human action recognition," *Image Vision Comput.* **28**(6), 976–990 (2010).
3. I. Haritaoglu, D. Harwood, and L. S. Davis, "W-4: real-time surveillance of people and their activities," *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(8), 809–830 (2000).
4. H.-I. Suk, B.-K. Sin, and S.-W. Lee, "Hand gesture recognition based on dynamic Bayesian network framework," *Pattern Recognit.* **43**(9), 3059–3072 (2010).
5. J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: a review," *ACM Comput. Surv.* **43**(3), (2011).
6. 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).
7. R. Chaudhry et al., "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1932–1939, IEEE Computer Society, Miami Beach, Florida (2009).
8. M. A. R. Ahad et al., "Motion history image: its variants and applications," *Mach. Vision Appl.* **23**(2), 255–281 (2012).
9. N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *European Conf. on Computer Vision*, pp. 428–441, Springer Berlin Heidelberg, Graz, Austria (2006).
10. R. V. Babu and K. R. Ramakrishnan, "Recognition of human actions using motion history information extracted from the compressed video," *Image Vision Comput.* **22**(8), 597–607 (2004).
11. M. Sizintsev and R. P. Wildes, "Spacetime stereo and 3D flow via binocular spatiotemporal orientation analysis," *IEEE Trans. Pattern Anal. Mach. Intell.* **36**(11), 2241–2254 (2014).
12. E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion," *J. Opt. Soc. Am. A* **2**(2), 284–299 (1985).
13. C. L. Huang and Y. T. Chen, "Motion estimation method using a 3D steerable filter," *Image Vision. Comput.* **13**(1), 21–32 (1995).
14. K. G. Derpanis et al., "Dynamic scene understanding: the role of orientation features in space and time in scene classification," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1306–1313, Providence, Rhode Island (2012).
15. L.-J. Li et al., "Object bank: an object-level image representation for high-level visual recognition," *Int. J. Comput. Vision* **107**(1), 20–39 (2014).
16. K. G. Derpanis et al., "Action spotting and recognition based on a spatiotemporal orientation analysis," *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(3), 527–540 (2013).
17. H. Zhou, L. Wang, and D. Suter, "Human action recognition by feature-reduced Gaussian process classification," *Pattern Recognit. Lett.* **30**(12), 1059–1066 (2009).
18. D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Comput. Vision Image Underst.* **115**(2), 224–241 (2011).
19. R. Chaudhry, G. Hager, and R. Vidal, "Dynamic template tracking and recognition," *Int. J. Comput. Vision* **105**(1), 19–48 (2013).
20. J. Dou and J. Li, "Robust human action recognition based on spatio-temporal descriptors and motion temporal templates," *Optik–Int. J. Light Electron Opt.* **125**(7), 1891–1896 (2014).
21. S. H. Jong and P. Milanfar, "Action recognition from one example," *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(5), 867–882 (2011).
22. V. Kellokumpu, G. Zhao, and M. Pietikainen, "Recognition of human actions using texture descriptors," *Mach. Vision Appl.* **22**(5), 767–780 (2011).
23. A. A. Efros et al., "Recognizing action at a distance," in *9th IEEE Int. Conf. on Computer Vision*, pp. 726–733, IEEE, Nice, France (2003).
24. E. Shechtman and M. Irani, "Space-time behavior-based correlation - OR - How to tell if two underlying motion fields are similar without computing them?," *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(11), 2045–2056 (2007).
25. W. He, K. C. Yow, and Y. Guo, "Recognition of human activities using a multiclass relevance vector machine," *Opt. Eng.* **51**(1), 017202 (2012).
26. J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vision* **79**(3), 299–318 (2008).
27. C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *17th Int. Conf. on Pattern Recognition*, pp. 32–36, IEEE Computer Society, Cambridge, England (2004).
28. B. Yogameena et al., "Human behavior classification using multi-class relevance vector machine," *J. Comput. Sci.* **6**(9), 1021–1026 (2010).
29. A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 2046–2053, IEEE Computer Society, San Francisco, California (2010).
30. W. T. Freeman and E. H. Adelson, "The design and use of steerable filters," *IEEE Trans. Pattern Anal. Mach. Intell.* **13**(9), 891–906 (1991).
31. K. G. Derpanis and J. M. Gryn, "Three-dimensional nth derivative of Gaussian separable steerable filters," in *Int. Conf. on Image Processing*, pp. 2777–2780, IEEE, Genoa, Italy (2005).
32. O. Michailovich, Y. Rathi, and A. Tannenbaum, "Image segmentation using active contours driven by the Bhattacharyya gradient flow," *IEEE Trans. Image Process.* **16**(11), 2787–2801 (2007).
33. Y. Rubner et al., "Empirical evaluation of dissimilarity measures for color and texture," *Comput. Vision Image Underst.* **84**(1), 25–43 (2001).
34. K. Grauman and T. Darrell, "The pyramid match kernel: efficient learning with sets of features," *J. Mach. Learn. Res.* **8**(Apr), 725–760 (2007).
35. L.-J. Li et al., "Object bank: an object-level image representation for high-level visual recognition," *Int. J. Comput. Vision* **107**(1), 20–39 (2014).
36. K. He et al., "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Euro. Conf. Comput. Vision*, pp. 346–361, Springer International Publishing, Zurich, Switzerland (2014).

37. L. Gorelick et al., "Actions as space-time shapes," *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(12), 2247–2253 (2007).
38. C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Trans. Intell. Syst. Technol.* **27**(3), 21–27 (2011).
39. T. Guha and R. K. Ward, "Learning sparse representations for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(8), 1576–1588 (2012).
40. A. Bouziane et al., "Unified framework for human behaviour recognition: an approach using 3D Zernike moments," *Neurocomputing* **100**, 107–116 (2013).
41. K. Huang, Y. Zhang, and T. Tan, "A discriminative model of motion and cross ratio for view-invariant action recognition," *IEEE Trans. Image Process.* **21**(4), 2187–2197 (2012).
42. W. Xinxiao et al., "Action recognition using context and appearance distribution features,"in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 489–496 (2011).
43. H. Ning et al., "Hierarchical space-time model enabling efficient search for human actions," *IEEE Trans. Circuits Syst. Video Technol.* **19**(6), 808–820 (2009).
44. G. Baudat and F. E. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Comput.* **12**(10), 2385–2404 (2000).
45. L. J. Cao et al., "A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine," *Neurocomputing* **55**(1–2), 321–336 (2003).
46. C. J. C. Burges, "Dimension reduction: a guided tour," *FNT Mach. Learn.* **2**(4), 275–365 (2010).

Yanhua Shao received an MS degree in pattern recognition and intelligent system from Southwest University of Science and Technology, China, in 2010. He is currently working toward a PhD in instrument science and technology at Chongqing University. His current research interest is in machine learning and action recognition.

Biographies for the other authors are not available.