Multimed Tools Appl
DOI 10.1007/s11042-011-0748-7

Exploring probabilistic localized video representation for human action recognition

Yan Song · Sheng Tang · Yan-Tao Zheng · Tat-Seng Chua · Yongdong Zhang · Shouxun Lin

Abstract In recent years, the bag-of-words (BoW) video representations have achieved promising results in human action recognition in videos. By vector quantizing local spatial temporal (ST) features, the BoW video representation brings in simplicity and efficiency, but limitations too. First, the discretization of feature space in BoW inevitably results in ambiguity and information loss in video representation. Second, there exists no universal codebook for BoW representation. The codebook needs to be re-built when video corpus is changed. To tackle these issues, this paper explores a localized, continuous and probabilistic video representation. Specifically, the proposed representation encodes the visual and motion information of an ensemble of local ST features of a video into a distribution estimated by a generative probabilistic model. Furthermore, the probabilistic video representation naturally gives rise to an information-theoretic distance metric of videos. This makes the representation readily applicable to most discriminative classifiers,

Y. Song · S. Tang · Y. Zhang · S. Lin
Laboratory of Advanced Computing Research, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 10090, China

S. Tang
e-mail: ts@ict.ac.cn

Y. Zhang
e-mail: zhyd@ict.ac.cn

S. Lin
e-mail: sxlin@ict.ac.cn

Y. Song (✉)
Graduate University of the Chinese Academy of Sciences, Beijing 10039, China
e-mail: songyan@ict.ac.cn

Y.-T. Zheng
Institute for Infocomm Research, A*STAR, Singapore, Singapore
e-mail: yzheng@i2r.a-star.edu.sg

T.-S. Chua
School of Computing, National University of Singapore, Singapore, Singapore
e-mail: chuats@comp.nus.edu.sg

Published online: 11 February 2011

such as the nearest neighbor schemes and the kernel based classifiers. Experiments on two datasets, KTH and UCF sports, show that the proposed approach could deliver promising results.

Keywords Human action recognition · Probabilistic video representation · Information-theoretic video matching

1 Introduction

Human action recognition in videos has spurred much research attention, as it has profound significance in building various multimedia applications, such as the unusual event detection in surveillance video and event detection in sports videos and movies, and so on. To date, human action recognition remains a challenging task, due to the huge variations in kinetic patterns of human movement, and photometric/geometric changes in subject appearance.

Inspired by the relative success of local features in image related applications [23], researchers recently shifted their focus to the local spatial temporal (ST) features for human action recognition [7, 17, 24, 28]. Local ST features represent videos in a compact but discriminative manner by describing local cuboids at the most informative spatial temporal locations. As shown in Fig. 1, spatial temporal interest points are detected in intensively moving locations. Despite of its good performance, the use of local ST features has limitations. The video representation is usually an ensemble of ST feature vectors. Thereby most discriminative classifiers cannot be used directly as most similarity metrics cannot be applied on variable length data (videos with different numbers of ST features).

One common solution is the vector quantization of local features, namely the bag of words (BoW) representation [7, 17, 28]. BoW representation in computer vision stems from the idea in the natural language processing domain that represents a document by the key words whose orders are ignored. In computer vision, an image or a video can be considered as a document and the features extracted from it are treated as "words". Specifically, it constructs a vocabulary by vector-quantizing the feature space and characterizing a video with the occurrences of each word in the vocabulary. A vocabulary is a prototype set in which each element is a representative of ST features. Then each video is represented by a histogram (BoW) of the vocabulary elements. Though BoW representation has shown good performance, it suffers from the following drawbacks. First, BoW scheme partitions the local feature space into discrete parts, in which each part corresponds to a visual word. Unlike textual document comprising discrete words, visual information is continuous in nature. The discrete partition inevitably brings in ambiguity, uncertainty and information loss in video representation. Second, the vocabulary is usually built on a subset of existing data. This makes the representation biased towards the videos whose features are involved in the vocabulary generation. A new vocabulary needs to be re-trained when it is applied on a new database or a new category is added to the existing database. However, adding new categories is common when the dataset is open for augmentation and the set of categories is not fixed beforehand.

Facing the limitations of the BoW representation, this paper explores a continuous, probabilistic and localized video representation scheme, based on and by expanding our previous work [30]. This representation models the visual and motion information embedded in the set of ST features of a video in an efficient and compact manner. We model each video as a distribution of localized ST features which is estimated by a

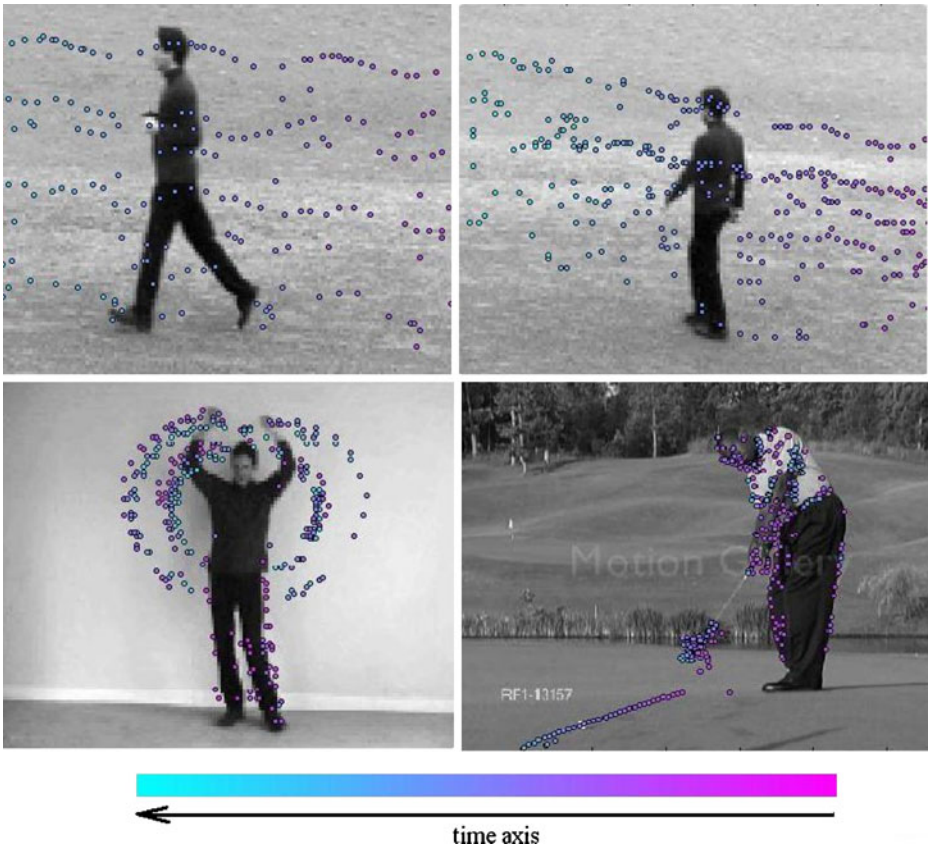

Fig. 1 Examples of local ST interest points detected in video sequence. The color of interest point indicates its temporal order

generative probabilistic model, such as the Gaussian Mixture Model (GMM) [2]. This probabilistic video representation naturally gives rise to information-theoretic distance metrics of videos. The distance measure of two videos with any lengths becomes the distance measure of two probabilistic distributions. The proposed probabilistic representation has several appealing properties. First, it takes into account the fact that human motion pattern is continuously distributed. In contrast to the BoW approach, it avoids the uncertainty caused in vector quantization of local ST features. Second, unlike BoW representing the set of ST features into a histogram, the proposed representation attempts to reveal the probabilistic structures of local ST features, as it takes full advantage of probabilistic generative models. More importantly, the proposed localized representation does not require any universal vocabulary. It considers a video as a probabilistic distribution in the ST feature space by estimating the probabilistic model parameters based on the set of ST features in the video. The representation is solely built on the ST features of individual videos independently. Compared to the BoW representation, it has more scalability by enabling existing action recognition systems to be readily applied on new video databases. Based on the probabilistic video representation and distance metric, most discriminative classifiers, such as the nearest neighbor schemes and the kernel based classifiers, are readily applicable for human action recognition.

In summary, the main contribution of this paper is that we explore a localized, continuous and probabilistic video representation that encodes the visual and motion information of an ensemble of local features. Testing on two datasets, i.e., KTH and UCF sports, shows that the proposed approach delivers promising results and outperforms the BoW approach with considerable margin.

The rest of the paper is organized as follows. In Section 2, related works about human action recognition and probabilistic representation are discussed. In Section 3, details of the distribution representation method and distance measure are represented. In Section 4, we demonstrate experiment results on two public datasets and Section 5 concludes the paper.

2 Related work

In early years, the main research efforts in human action recognition focused on the tracking of human subject and the extraction of holistic features like shape of gesture in each frame [4, 32]. Veeraraghavan et al. [32] adopted a statistical shape representation [13] method to describe the shape of human body. They described human body silhouette as a k-dimensional complex vector where k is the number of landmarks on the shape. Davis and Bobick [4] proposed a temporal template based representation which is a static vector image with each point setting by the value of a function of the motion properties at the corresponding spatial location. It was based on the assumption that the motion of the object can be separated. Despite of the simplicity, these methods suffer from two drawbacks: (1) the performance relies highly on tracking and segmentation, which are still open research problems; and (2) they are sensitive to occlusion and cluttered background.

Motivated by the development of local features in image classification and object categorization [23], researchers shifted their attention to local ST feature based models for human action recognition [7, 14, 16, 17, 29, 35]. Doll'ar et al. [7] applied filters on spatial and temporal domains and extracted cuboids at locations of maximum response. Gradients of each pixel in a cuboid are concatenated into a vector to describe the cuboid. Kl¨aserv et al. [14] proposed a 3D extension of the famous SIFT descriptor [23]. It generated histograms of 3D gradient orientation base on integral video representation. Similarly, Scovanner et al. also developed a 3D SIFT alike feature [29]. Laptev and Lindeberg extended the Harris detector [16]. The histograms of spatial gradient and optical flow were computed to generate HOG/HOF descriptor [17]. Willems et al. developed an extended version of Hessian saliency measure to locate ST interest points [35].

Based on the aforementioned local ST features, many human action recognition methods have been explored. Researchers explored some improvements for the vocabulary generation for the BoW representation. Liu and Shah [21] proposed to automatically discover the optimal size of vocabulary by utilizing the principle of maximization of mutual information (MMI). Ballen et al. [1] adopted a radius-based clustering method and a soft assignment to construct a codebook. In addition, Niebles et al. [24] proposed an unsupervised learning method by learning the probability distributions of ST words and the intermediate topics corresponding to human action categories via probabilistic Latent Semantic Analysis (pLSA) [12] and Latent Dirichlet Allocation (LDA) [3]. Laptev et al. [17] addressed the importance of human action recognition in realistic videos. They presented a method combining and extending several ideas including local ST features, ST pyramids and multi-channel non-linear SVM classifiers. Liu et al. [22] proposed an approach for generic visual vocabulary. They adopted diffusion maps to learn a semantic visual vocabulary on quantized midlevel features which was represented by the vector of

mutual information. The idea was to embed the midlevel feature into a semantic lower-dimensional space to construct a semantic visual vocabulary. All the aforementioned approaches share one commonality: a discrete BoW scheme is used as video representation. Facing the ambiguity and information loss in BoW, we argue that a probabilistic and continuous representation is more suited for modeling human motion.

Probabilistic representation of data has been widely used in various disciplines, such as speaker identification in audio signal processing domain [38] [37], and so on. In computer vision community, image or video was represented by probabilistic distribution of pixel [8, 9] or sub-window features [31]. One important aspect of probabilistic data representation is distance metric. Kullback–Leibler (KL) measure [15] has been applied in image matching task, together with Gaussian Mixture model [8]. Do and Vetterli [6] adopted generalized Gaussian Density and KL-distance for texture retrieval. Vasconcelos and Moreno [31] investigated the advantage of KL-kernel to combine discriminant recognition with representation for visual recognition. Cao et al. [18] employed a GMM to represent universal background distribution and an adaptation GMM for action model. Our proposed representation is similar to the approaches above; in the way they all summarize information in a probabilistic process. However, different from the above distribution-based audio, image and video modeling methods, our work views action videos as a 3D volume represented by an ensemble of local ST features which encodes the most informative parts for action recognition. Moreover, the proposed representation tackles information localization too. Namely, the probabilistic representation needs to preserve the robustness to clutter and occlusions in local ST features.

In this aspect, our method is similar to the work by Zhou et al. [39] in part. The approach described the bag of SIFT features in the video frames by a specialized GMM adapted from a global GMM. However, our method is different from it in three-fold. First, our method adopts local ST feature which has been proved effective in many previous works [7, 24, 28] due to its property of encoding not only static but also motion information in videos which is crucial to action recognition. Zhou et al. [39] adopted SIFT feature [23] for video representation to analyze event. Second, our GMM based representation avoids dependence on any global information. One of the most important characteristics of our approach is that it is a localized representation for videos, which means each video is independently represented. Zhou et al. [39] firstly estimated a global GMM from the whole dataset and then adapted a specialized GMM for each video. Third, our approach exploits the Minimum Description Length (MDL) criterion to determine number of GMM components automatically. This ensures that our probabilistic description of each video is data-driven without any prior knowledge and assumption. Our previous work proposed a distribution based representation of videos for human action recognition [30]. This work extends the distance metrics and algorithms of our previous work in two-fold. First, we explore the information-theoretic distance metrics for distributions to test the sensitivity of the proposed representation against different distance metrics. Second, we conduct more extensive experiments to verify the effectiveness of the proposed method from different angles and add the benchmark with several state-of-the-art methods.

3 Algorithm

Figure 2 shows the overall framework of our system. As shown, the approach first extracts ST features from videos and learn the probabilistic video representation from ST features via a generative probabilistic model. Information-theoretic distance metrics are then

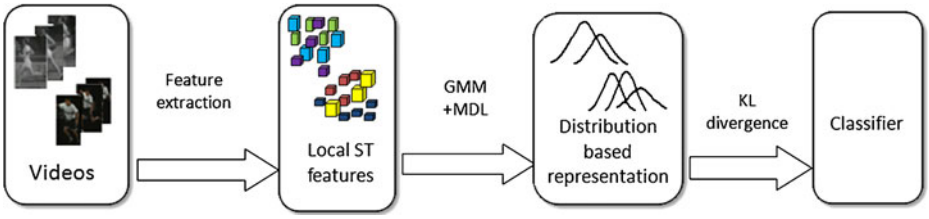

Fig. 2 Flowchart of the proposed system

exploited as similarity measure. Finally, a discriminative classifier is applied to perform action recognition. For clarity, we list notations of variables used in the method in Table 1.

3.1 Local ST Feature extraction

Motivated by the promising result and the denseness of interest points [24] [22], we generate ST features by adopting the method proposed by Doll'ar et al. [7]. The procedure of feature extraction includes four steps.

Firstly, a video is considered as a 3D volume *I(x, y, t)* that a Gaussian filter and two Gabor filters are applied on the spatial and temporal domains respectively. The response function of the filters is defined as:

$$R(x,y,t) = [I * g_\sigma(x,y) * h_{ev}(t)]^2 + [I * g_\sigma(x,y) * h_{od}(t)]^2, \quad (1)$$

Table 1 Notation of symbols

Symbol	Description
g_σ (x,y)	2D Gaussian filter
h_{ev} $(t)/h_{od}$ (t)	A quadrature pair of Gabor filters
R	Response function
σ	Parameter in Gaussian filter
τ	Parameter in Gabor filters
Δ	Size of cuboid
K	The number of components in a GMM
α_k	The weight of the k^{th} Gaussian component
μ_k	The mean of the k^{th} Gaussian component
Σ_k	Covariance matrix of the k^{th} Gaussian component
F	ST feature set
$\gamma^{(n)}(i,k)$	The probability that a local ST feauture f_i is generated by the k^{th} Gaussian component in the n^{th} iteration
N	The number of features
M	The dimension of the feature vector
$KLD(G_1\|\|G_2)$	KL divergence of distributions G_1 and G_2
$KLD(G_1, G_2)$	Symmetric KL divergence of distributions G_1 and G_2
$JSD(G_1\|\| G_2)$	JS divergence of distributions G_1 and G_2
$H(P)$	Shannon entropy of the distribution P
$KLD_{var}(G_1, G_2)$	Variational lower bound approximation of KL divergence of distributions G_1 and G_2
$JSD_{var}(G_1, G_2)$	Variational lower bound approximation of JS divergence of distributions G_1 and G_2

where $g_\sigma(x,y)$ is the 2D Gaussian filter applied on spatial dimension xy, and $h_{ev}(t)$ and $h_{od}(t)$ are a quadrature pair of Gabor filters applied on temporal dimension t. The Gabor filters are defined as $h_{ev}(t;\tau,\omega) = -\cos(2\pi t\omega)\exp(-t^2/\tau^2)$ and $h_{od}(t;\tau,\omega) = -\sin(2\pi t\omega)\exp(-t^2/\tau^2)$ with $\omega = 4/\tau$. The parameters σ and τ can be considered as spatial and temporal detector scales respectively.

Secondly, interest points are located at the local maximums of the response R. Intensive spatial-temporal changes occur at these locations. Specifically, the local maximums correspond to locations with significant spatial characteristics and complex motion.

Thirdly, local 3D cuboid of the interest point (x,y,t) is extracted with the size of $\Delta_x(\sigma) * \Delta_y(\sigma) * \Delta_t(\tau)$, where $\Delta_x(\sigma) = \Delta_y(\sigma) = 2*ceil(3\sigma) + 1$ and $\Delta_t(\tau) = 2*ceil(3\tau) + 1$. A cuboid contains windowed neighbor pixels of the interest point.

At last, each cuboid is characterized by vectors concatenated by brightness gradients of pixels in them. The feature dimension is reduced by principal component analysis (PCA).

3.2 Probabilistic and localized representation of videos

In human action videos, local ST interest points are located at the most intensively changing parts and extracted features characterize the corresponding local motion patterns. We model a video as a distribution of localized ST features. There exist several generative probabilistic models that can learn the distributions, like GMM and Hidden Markov Models (HMM) [25]. HMM models a system as a Markov process which is a random process. A random process amounts to a sequence of random variables known as a temporal or spatial series. ST interest points extracted from a video distribute in a 3-D space and they cannot be directly ordered one by one as a time series. Thereby we cannot directly utilize HMM in this situation. Here, we choose GMM due to its simplicity and good practical performance [6, 8, 9, 31, 37, 38]. Let $F = \{f_i |\ i = 1, 2, \ldots, N\}$ denotes the local ST feature set extracted from a video as described in Section 3.1, where N denotes the number of ST features in this video. Parameters of a GMM are denoted as $\{K, \theta\}$ and $\theta = \{\alpha_k, \mu_k, \Sigma_k |\ k = 1, 2\ldots, K\}$, where K denotes the number of components, α_k, μ_k and Σ_k denote the weight, mean and covariance matrix respectively. The log likelihood of local ST feature set in a video is given by:

$$\log p(\mathrm{F}|K,\theta) = \sum_{i=1}^{N} \log\Big(\sum_{k=1}^{K} \alpha_k p(f_i|k,\theta)\Big). \tag{2}$$

The probability $\gamma(i,k)$ that a local ST feauture f_i is generated by the k^{th} Gaussian component is computed by:

$$\gamma(i,k) = \frac{\alpha_k p(f_i;\mu_k,\Sigma_k)}{\sum_{k'=1}^{K} \alpha_{k'} p(f_i;\mu_{k'},\Sigma_{k'})}, \tag{3}$$

$$p(f;\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}\det(\Sigma)^{1/2}} \exp\Big\{-\frac{1}{2}(f-\mu)^T \Sigma^{-1}(f-\mu)\Big\}. \tag{4}$$

We estimate GMM by Expectation-Maximization (EM) algorithm. In the n^{th} iteration of EM algorithm, parameters of the k^{th} Gaussian model and the mixture weight are computed

by the expectation value of hidden variable γ obtained in the $(n-1)^{\text{th}}$ iteration, as defined by Eqs. 5, 6 and 7 below:

$$\alpha_k^{(n)} = \frac{1}{N} \sum_{i=1}^{N} \gamma^{(n-1)}(i,k), \tag{5}$$

$$\mu_k^{(n)} = \frac{1}{\sum_{i=1}^{N} \gamma^{(n-1)}(i,k)} \sum_{i=1}^{N} \gamma^{(n-1)}(i,k) f_i, \tag{6}$$

$$\Sigma_k^{(n)} = \frac{1}{\sum_{i=1}^{N} \gamma^{(n-1)}(i,k)} \sum_{i=1}^{N} \gamma^{(n-1)}(i,k)(f_i - \mu_k^{(n)})(f_i - \mu_k^{(n)})^T. \tag{7}$$

Then the expectation value of hidden variable γ is updated by the parameters of the Gaussian model and the mixture weight in the n^{th} iteration:

$$\gamma^{(n)}(i,k) = \frac{\alpha_k^{(n)} p(f_i; \mu_k^{(n)}, \Sigma_k^{(n)})}{\sum_{k'=1}^{K} \alpha_{k'}^{(n)} p(f_i; \mu_{k'}^{(n)}, \Sigma_{k'}^{(n)})}. \tag{8}$$

The iteration is terminated when the convergence criterion is met. Convergence of the algorithm can be determined by observing the change of log-likelihood of the data.

After modeling an action video by a GMM, the video is represented by the parameters of a GMM *{K, θ}*. In another view, we can cluster the ST features in a video corresponding to different Gaussian components. Thereby, the GMM based method represents the local feature set by comprehensively characterizing several clusters in a distribution manner. We give some visualization examples of the clustering effect by GMM modeling in the experiment section.

There is a parameter in the probabilistic representation that should be decided before EM algorithm which is the number of components in GMM. However, in human action recognition, the number of Gaussian mixtures should be determined without any prior knowledge. As we know, MDL criterion is a widely used method in model selection. Here, we utilize the MDL criterion to choose the optimal number of components in our probabilistic representation for action videos. The MDL criterion adds a penalty term to the log likelihood to prevent over-fitting by implementing a tradeoff between the complexity of the hypothesis and the complexity of the data given the hypothesis. The MDL criterion is to minimize the following function:

$$MDL(K,\theta) = -\log p(f|K,\theta) + \frac{1}{2} L \log(NM), \tag{9}$$

where N is the number of features, and M is the dimension of the feature vector. L is given by:

$$\mathrm{L} = K\left(1 + M + \frac{(M+1)M}{2}\right) - 1 \tag{10}$$

Although we have the object function, we cannot obtain the optimal solution directly. Here, we initialize K by a large number and decrease it one by one. For each K, the penalty

term is fixed so that we only need to apply EM iteration to minimize MDL. To decrease K, we merge the nearest two mixtures by the following equations:

$$\alpha_{(a,b)} = \alpha_a + \alpha_b, \tag{11}$$

$$\mu_{(a,b)} = \frac{\alpha_a \mu_a + \alpha_b \mu_b}{\mu_a + \mu_b}, \tag{12}$$

$$\Sigma_{(a,b)} = \frac{\alpha_a(\Sigma_a + (\mu_a - \mu_{(a,b)})(\mu_a - \mu_{(a,b)})^t)}{\mu_a + \mu_b} + \frac{\alpha_b(\Sigma_b + (\mu_b - \mu_{(a,b)})(\mu_b - \mu_{(a,b)})^t)}{\mu_a + \mu_b}. \tag{13}$$

Once we obtain the parameters of the new merged mixture, they are set to be the initial parameters of the new EM iteration for $K-1$. After an EM iteration for a K, we can compute the MDL by Eq. 9. We choose the K with the minimum MDL to be the optimal number of mixtures in the GMM.

Now we can model a local ST feature set by a GMM with an automatically chosen number of components for each action video.

3.3 Information-theoretic distance metrics of videos

In our framework, a video is represented by an ensemble of local ST features, which is further modeled by a probability distribution. The problem of distance metric of videos then becomes the issue of distance metric between probability density functions (PDF).

Existing distance measures for multiple probability distributions, such as the KL measure and the Jensen-Shannon (JS) divergence [19], have been used as distance and kernel in some applications [5, 33]. Here we adopt and compare these two divergences. The KL divergence is commonly defined as:

$$KLD(G_1||G_2) = \int g_1(x) \log \frac{g_1(x)}{g_2(x)} dx. \tag{14}$$

We adopt symmetric KL divergence by adding two terms:

$$\begin{aligned} \mathrm{KLD}(G_1, G_2) &= KLD(G_1||G_2) + KLD(G_2||G_1) \\ &= \int g_1(x) \log \frac{g_1(x)}{g_2(x)} dx + \int g_2(x) \log \frac{g_2(x)}{g_1(x)} dx. \end{aligned} \tag{15}$$

The JS divergence of n probability distributions $\{P_i|\ i=1,2,...n\}$ is defined as:

$$JSD(P_1, P_2, ..., P_n) = H(\sum_{i=1}^{n} \pi_i P_i) - \sum_{i=1}^{n} \pi_i H(P_i), \tag{16}$$

where $\pi = \{\pi_1, \pi_2, ..., \pi_n | \pi_i > 0, \sum \pi_i = 1\}$ are the weights of distributions $\{P_i|\ i=1,2,...n\}$ and $H(P)$ is the Shannon entropy of the distribution P formulated as:

$$H(P) = -\int_{\Omega} p(x) \log p(x) dx. \tag{17}$$

For two distributions case described above, the JS divergence is:

$$JSD(G_1||G_2) = H[wg_1(x) + (1-w)g_2(x)] - wH[g_1(x)] - (1-w)H[g_2(x)]. \quad (18)$$

By substituting for H and setting w=1/2, JSD is a symmetrized and smoothed version of the KL divergence $KLD\ (G_1||G_2)$ by:

$$JSD(G_1||G_2) = \frac{1}{2}KLD(G_1||S) + \frac{1}{2}KLD(G_2||S), \quad (19)$$

$$s(x) = \frac{1}{2}g_1(x) + \frac{1}{2}g_2(x), \quad (20)$$

Unfortunately, there is no closed form expression for the KL divergence between two GMMs and it is usually done by Monte-Carlo simulations. To compute Monte-Carlo simulation, we need to draw samples from PDF in order to get the expectation of $log\ (G_1/G_2)$. Due to the high computation complexity for processing video data, we adopt the variational lower bound [11] to compute an approximation of KL divergence of two GMMs G_1 and G_2 by:

$$KLD_{\text{var}}(G_1||G_2) = \sum_a \alpha_a \log \frac{\sum_{a'} \alpha_{a'} e^{-KLD(G_{1a}||G_{1a'})}}{\sum_b \alpha_b e^{-KLD(G_{1a}||G_{2b})}}, \quad (21)$$

where $KLD_{var}(G_1||G_2)$ is the variational lower bound approximation of KL divergence; $KLD\ (G_{1a}||G_{1a'})$ denotes the KL divergence between two Gaussian components of GMM G_1; and $KLD(G_{1a}||G_{2b})$ denotes the KL divergence between one Gaussian component of GMM G_1 and one Gaussian component of GMM G_2. The symmetric KL divergence of two GMMs is approximated as:

$$\begin{aligned} \text{KL}D_{\text{var}}(G_1, G_2) = &\sum_a \alpha_a \log \frac{\sum_{a'} \alpha_{a'} e^{-KLD(G_{1a}||G_{1a'})}}{\sum_b \alpha_b e^{-KLD(G_{1a}||G_{2b})}} \\ &+ \sum_b \alpha_b \log \frac{\sum_{b'} \alpha_{b'} e^{-KLD(G_{2b}||G_{2b'})}}{\sum_a \alpha_a e^{-KLD(G_{2b}||G_{1a})}}. \end{aligned} \quad (22)$$

And the JS divergence of two distributions G_1 and G_2 is approximated as:

$$\begin{aligned} JSD_{\text{var}}(G_1||G_2) = &\frac{1}{2}\sum_a \alpha_a \log \frac{\sum_{a'} \alpha_{a'} e^{-KLD(G_{1a}||G_{1a'})}}{\frac{1}{2}\sum_{a'} \alpha_{a'} e^{-KLD(G_{1a}||G_{1a'})} + \frac{1}{2}\sum_b \alpha_b e^{-KLD(G_{1a}||G_{2b})}} \\ &+ \frac{1}{2}\sum_b \alpha_b \log \frac{\sum_{b'} \alpha_{b'} e^{-KLD(G_{2b}||G_{2b'})}}{\frac{1}{2}\sum_{b'} \alpha_{b'} e^{-KLD(G_{2b}||G_{2b'})} + \frac{1}{2}\sum_a \alpha_a e^{-KLD(G_{2b}||G_{1a})}}. \end{aligned} \quad (23)$$

The KL divergence of two single Gaussians a and b is formulated as:

$$KLD(a||b) = \frac{1}{2}[\log \frac{|\Sigma_b|}{|\Sigma_a|} + tr(\Sigma_b{}^{-1}\Sigma_a) - d + (\mu_a - \mu_b)^t \Sigma_b{}^{-1}(\mu_a - \mu_b)]. \quad (24)$$

3.4 Analysis

Here we discuss the properties of the proposed probabilistic video representation in various aspects and compare it with BoW approach to demonstrate that the probabilistic video representation is more suitable for human action recognition task.

3.4.1 Continuous encoding of visual-motion information

The proposed probabilistic representation characterizes the visual and motion information in a localized and continuous manner. This appealing property makes it intrinsically fit for the encoding of local ST-features, as local ST-features are continuous-valued variables, in a statistical perspective. On the other hand, the bag-of-words approach is a lossy data encoding scheme based on the principle of block coding. The vector quantization in bag-of-words approach divides the feature space into discrete partitions. For example, the feature space is partitioned into k subspace which is decided by the cluster centers by adopting k-means clustering. This discrete partitioning of continuous feature space inevitably brings in information loss, and further, ambiguity and uncertainty in video representation. For example, the local features in a cluster/partition do not necessarily carry similar visual-motion information, as vector quantization is always a local optimum that assigns feature spaces to different partitions in a tradeoff manner.

3.4.2 No codebook is required

The proposed probabilistic video representation requires no codebook to generate beforehand. The encoding of visual motion information relies solely on the local ST features in a video. This not only provides better efficiency, but also enables the representation to generalize to different datasets and tasks better.

In contrast, the bag-of-words approach has a visual-motion codebook built on the existing training data. This makes the representation biased towards the videos whose features are involved in the vocabulary generation. A new vocabulary has to be re-trained when it is applied on a new database or a new category is added to the existing database. For example, if a new action category is added to the dataset after a vocabulary has been generated on the existing training set, the vocabulary may not suffice to represent the features appear in the new category. Adding new categories is common when the database is open for augmentation and the set of categories is not fixed beforehand. In this aspect, the proposed probabilistic representation method is better than the BoW method in extensibility and flexibility.

3.4.3 Minimum parameter tuning in video representation

The construction of the proposed video representation requires minimum parameter tuning. The number of mixture components in the representation is automatically determined by MDL criterion, in the formulation of model selection. In contrast, the number of video-words in bag-of-words approach is always an open question. To determine how coarsely or finely to quantize the local feature, researchers have to set the number of video-words empirically.

3.4.4 Distance metric

A distance metric function should obey three properties: isolation, symmetry and triangular inequality. KL divergence is not symmetric. In this work, we adopt symmetric KL distance to tackle this issue. Nor does it obey the triangular inequality. Hence, KL divergence is not a true distance metric. However, it generates a topology on the space of probabilistic distributions. Compared to KL measure, JS divergence is symmetric, well-defined, always a finite value and the square root of it is a metric. Both the two divergences are usually utilized to be the “distance metric” in’ probabilistic distribution space.

4 Experiments

4.1 Experimental setup

To test the proposed video representation, we employ two public datasets: KTH [28] and UCF sports datasets [20]. The KTH dataset is one of the most widely used datasets for human action recognition. It contains 6 action categories ("boxing", "hand clapping", "hand waving", "jogging", "running" and "walking") performed by 25 subjects in four scenes including "outdoors", "outdoors with different clothes", "outdoors with scale variation" and "indoors". There are 2,391 videos in the dataset in total; each video clip contains one person executing one action. The UCF sports dataset contains various sports videos from broadcast television in ten categories ("diving", "golf swinging", "kicking", "lifting", "riding horse", "running", "walking", "swinging angle", "swing bench" and "skate boarding"). This collection is in a wide range of scenes in unconstrained environment so that the intra-class variability is large. Hence, it is more challenging compared with the KTH dataset. There are 150 videos in the dataset. Figure 3 shows sample frames from these two datasets.

For the KTH dataset, we follow the original setup [28] to divided the dataset into the training set (eight people), validation set (eight people) and the test set (nine people). We use the average of recognition accuracy as the evaluation criteria. For the UCF sports dataset, we use half the data for training and the rest for testing, and run it ten times to report the average performance as done by Liu et al. [20] for benchmark purpose.

The scale parameters σ and τ in ST feature extraction are both set to 1.5. Compared with our previous work [30], we tune the scale parameters in feature extraction to obtain more dense interest points. The dimension of local ST feature is reduced to 30 by PCA for efficiency. We adopt two classifiers on both datasets including K-Nearest Neighbors (KNN) and Support Vector Machine (SVM). The regularization parameter C is set to 1000 in SVM. In the case of multi-class classification we adopt the one-against-all scheme. As SVM is kernel based classifier, we adopt the Gaussian kernel defined by:

$$K(X_1, X_2) = \exp(-\frac{1}{A} D(X_1, X_2)), \tag{25}$$

where $D(X_1,X_2)$ is the distance measure between two representations which is the JS divergence or the KL divergence for the proposed representation. A is a scaling parameter set to the mean value of the distances between all training samples [31].

Fig. 3 Sample keyframes from two datasets: **a** KTH, and **b** UCF sports

4.2 Experimental results

To demonstrate the effect of the probabilistic representation of videos, we visualize the distribution of GMM modeling of local spatial temporal interest points in 2D and 3D spaces. As shown in Fig. 4, different colors indicate different components of GMM which interest points belong to. Column (a) shows all of the interest points in the video stacked in

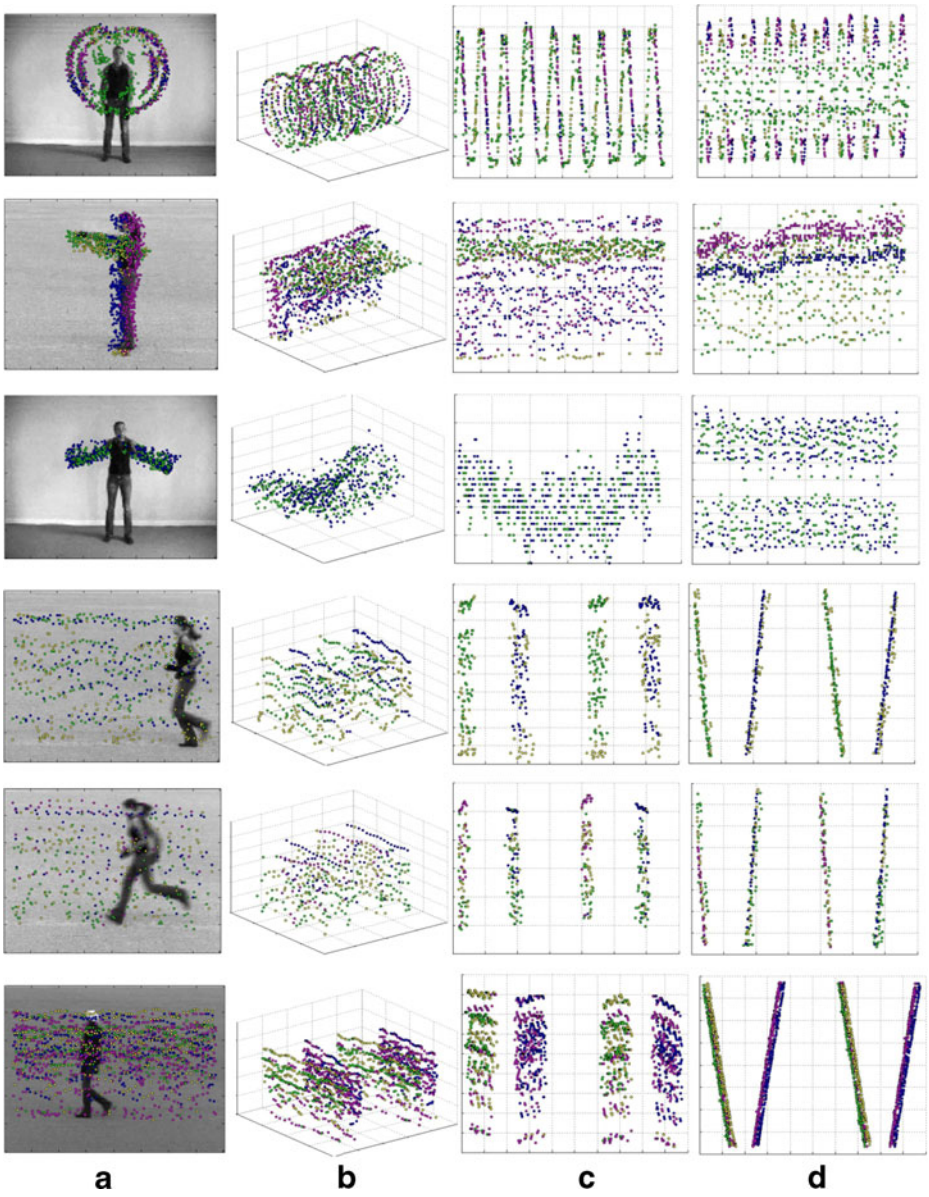

Fig. 4 Visualization of GMM modeling of local ST points in action videos. Column **a** shows all of the interest points in the video stacked in one frame. Column **b** shows the interest points in the original 3D space *(x,y,t)*. Columns **c** and **d** are the stereo images in column **b** projecting on *xt* and *yt* planes

one frame. Column (b) shows the interest points in the original 3D space *(x,y,t)*. Columns (c) and (d) are the stereo images in column (b) projecting on *xt* and *yt* planes. We notice that interest points belonging to the same component tend to gather around certain part of human body and correspond to certain direction of the action.

To assess the effectiveness of the proposed video representation, we evaluate three representations with same experimental setting as follows:

Rep-1: the BoW representation;
Rep-2: the fixed GMM representation with fixed number of Gaussian components;
Rep-3: the proposed video representation.

For the BoW representation, we sample a subset of 100 k features from all the features extracted from the training videos and obtain k (codebook size) feature prototypes by k-means clustering. χ^2 distance is adopted for distance metric:

$$\mathrm{D}(his_i, his_j) = \frac{1}{2}\sum_{b=1}^{B}\frac{[his_i(b) - his_j(b)]^2}{his_i(b) + his_j(b)} \tag{26}$$

where his_i and his_j are histograms of sample i and j and B is the number of bins in the histogram. For both GMM based representations, JS divergence (JSD) or KL divergence (KLD) is adopted for distance metric as introduced in Section 3.3

4.2.1 The KTH dataset

Firstly, we test the impact of some parameters to the system and compare the average recognition accuracies of three representations by 1-NN classifier on the KTH dataset. We tune the codebook size of the BoW representation from 100 to 1,000. As shown in Fig. 5a, it achieves the highest accuracy when codebook size is 500. The recognition accuracy is in the range from 81% to 84.5%. For both GMM based representations, i.e. Rep-2 and -3, we adopt JS divergence and KL divergence as distance metric. We tune the number of fixed

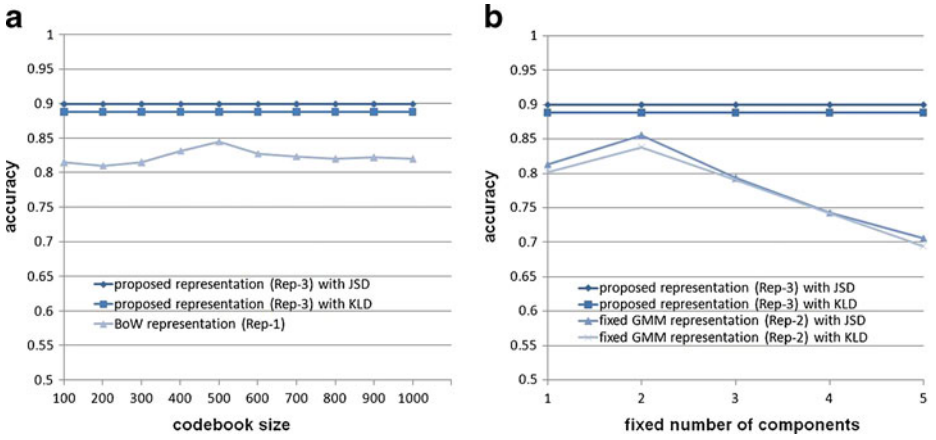

Fig. 5 **a** Average recognition accuracies by the proposed representation and the BoW representation with different codebook size on the KTH dataset with 1-NN classifier. **b** Average recognition accuracies by the proposed representation and the fixed GMM representation with different number of components of GMM on the KTH dataset with 1-NN classifier

GMM from one to five. As shown in Fig. 5b, the result of the fixed GMM representation with JSD is slightly better than that with KLD. It achieves the highest accuracy when the number of components is set to two. The recognition accuracy of the fixed GMM representation with JSD is in the range from 70.6% to 85.5%. The performance decreases fast when the number of components is larger than two. We conjecture the reason is that the actions as well as the shooting environment in the KTH dataset are relatively simple. Too many Gaussian components may bring in the problem of over-fitting. The proposed representation with JSD achieves the average accuracy of 90% which is better than the best performances of the other two representations. This manifests two points: (1) by encoding visual-motion information in a continuous and probabilistic manner, the proposed video representation can deliver better performance than the BoW approach; and (2) the number of mixture components in GMM plays an important role and our proposed MDL criterion based model selection gives good performance.

Then we test the performance of the three representations with different classifiers. For the first two representations, we choose the fixed GMM representation with two components in each GMM and set the codebook size to 500 for the BoW representation. This parameter setting has been verified to be the optimal for these two representations in the previous experiment. We adopt K-NN classifier with parameter K set from one to eight. We also adopt both JS divergence and KL divergence for the two distribution based representations. The results are shown in Fig. 6. For both representations (Rep-2 and -3), JS divergence performs a little better than the KL divergence. Compared with the first two representations, parameter variation in K-NN classifier has little effect on the performance of the proposed representation. The proposed representation performs best with the 6-NN classifier. It is noticed that the recognition accuracy of the proposed representation always outperforms the other two representations with considerable margin. The improvement

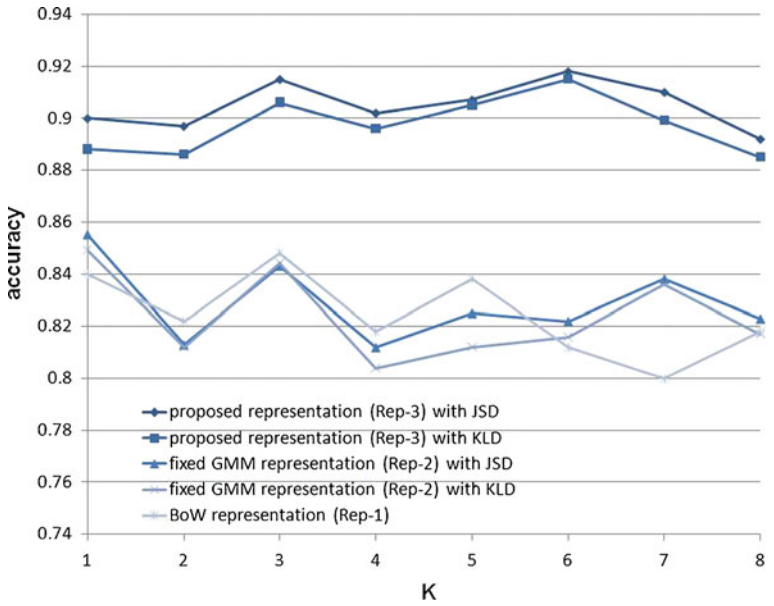

Fig. 6 Average recognition accuracies with different values of parameter K in the K-NN classifier for the fixed GMM representation, BoW representation and proposed representation on the KTH dataset

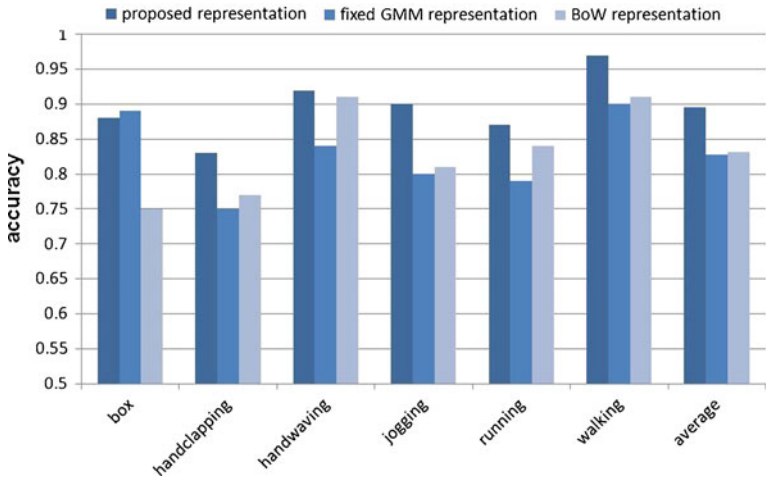

Fig. 7 Recognition accuracies of all action categories by the fixed GMM representation, the BoW representation and the proposed representation with SVM classifier on the KTH dataset

achieves to about 8% on average. We also notice that the proposed representation performs well with both the distance metrics.

We further examine the recognition performance of individual action categories. We adopt the SVM classifier. As JS divergence is observed to show a little advantage over KL divergence, we only use JS divergence as the distance metric for the distribution based representations in this experiment. Figure 7 shows the experiment results of recognition accuracies on all action categories in detail on the KTH dataset. For most categories the proposed representation performs best. Especially, the proposed representation surpasses the fixed GMM representation on "jogging" category by almost 10% and surpasses the BoW representation on "boxing" category by almost 13%. The average accuracies of the proposed representation, the fixed GMM representation and the BoW representation are 89.5%, 82.83% and 83.17% respectively. We notice that the advantage of the proposed representation is evident no matter which classifier is adopted.

Benchmark We also benchmark our approach with state-of-the-art methods [3–5, 9. 19, 20, 25] reported by other researchers. Kl¨aser et al. focused on a new spatial-temporal feature extraction which was based on histogram of oriented 3D gradients [14]. Laptev et al. adopted local ST feature combined with space-time pyramids and multi-channel SVMs to obtain the optimal descriptor and grid [17]. Ballan et al. proposed a 3D gradient descriptor and a radius-based clustering method to generate codebook [1]. Although these methods

Table 2 Benchmark on the KTH dataset

Method	Schuldt et al. [28]	Doll'ar et al. [7]	Niebles et al. [24]	Wong et al. [36]
Accuracy	0.717	0.817	0.833	0.866
Method	Kl¨aser et al. [14]	Laptev et al. [17]	Ballan et al. [1]	Our best
Accuracy	0.914	0.918	0.926	0.918

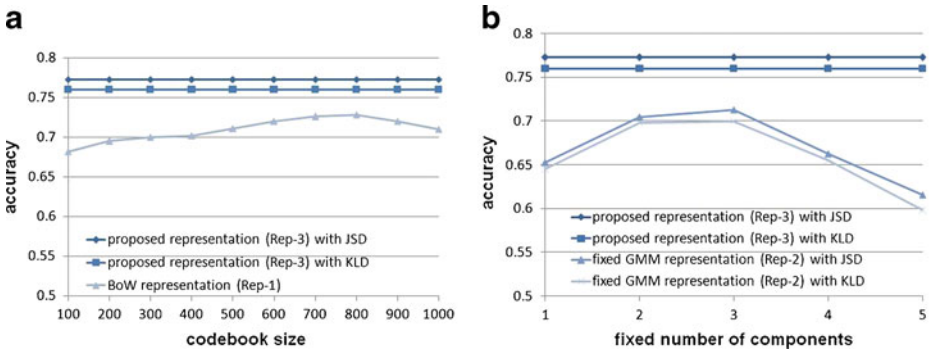

Fig. 8 **a** Average recognition accuracies by the proposed representation and the BoW representation with different codebook size on the UCF sports dataset with 1-NN classifier. **b** Average recognition accuracies by the proposed representation and the fixed GMM representation with different number of components of GMM on the UCF sports dataset with 1-NN classifier

adopted different local ST features extraction approaches and different recognition schemes, most of them were based on BoW representation. As shown in Table 2, our approach outperforms most of the existing works and is comparable to the best. We notice the method based on radius-based clustering [1] is appreciably better than ours. The good performance is partially because of the combination of the histogram concatenation of sub-region computed from 3D gradient and the histogram of optic flow. However, the focus of our work is not ST feature combination. Actually, their accuracy by only the 3D gradient is 90.38% [1]. We attribute the improvement of our representation over BoW representation to the continuous coding of local ST features, as the information loss is less than the discrete one of BoW.

4.2.2 The UCF sports dataset

To further validate the effectiveness of our approach, we perform similar experiments with the three representations on the UCF sports dataset. Firstly, we adopt 1-NN classifier and

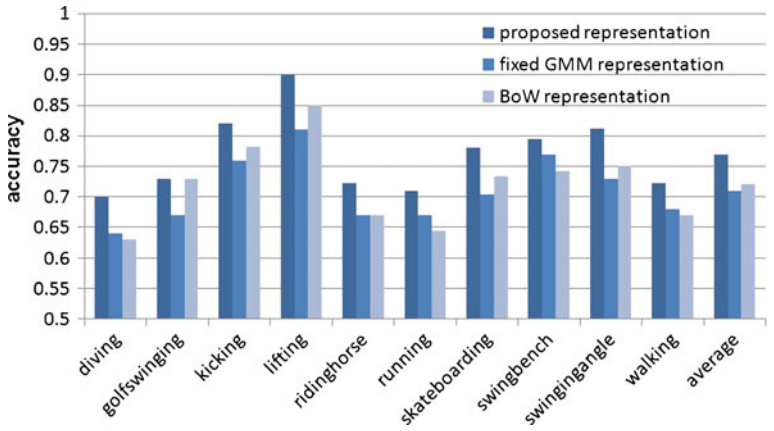

Fig. 9 Recognition accuracies of all action categories by the fixed GMM representation, the BoW representation and the proposed representation with SVM classifier on the UCF sports dataset

Table 3 Benchmark on the UCF sports

	Rodriguez et al. [27]	Liu et al. [20]	Wang et al. [34]	Ours	Ours with ROI
Accuracy (%)	69.2	79.6	76.6	77.3	81.1

observe the impact of some parameters of the first two representations (Rep-1 and -2). We tune the codebook size of the BoW representation from 100 to 1,000. It achieves the highest accuracy when codebook size is 800 as shown in Fig. 8a. The accuracy is in the range from 68.2% to 72.8%. For both GMM based representations, we adopt JS and KL divergence as distance metric. We also tune the number of the fixed GMM from one to five. It achieves the highest accuracy when the number of components is set to three as shown in Fig. 8b. The accuracy of the fixed GMM representation with JSD varies from 61.6% to 71.3%. The average accuracy of our proposed representation is 77.3% with JSD and 76% with KLD. Similar to the KTH dataset, our representation gives consistently better results than the other two representations.

Then we also test the performance of three representations with SVM classifier. For the first two representations (Rep-1 and -2), we choose the fixed GMM representation with three components in each GMM and set the codebook size to 800 for the BoW representation. Here, we adopt JSD for both the distribution based representations. As shown in Fig. 9, the proposed representation performs best on average accuracy with 76.9%, which outperforms the fixed GMM representation and the BoW representation by 5.9% and 4.8% respectively. The proposed representation outperforms the other two representations on most categories. As the UCF sports dataset is comprised of real action videos and is more complex, we argue that the proposed representation is effective and robust in real video situation. We draw similar conclusion with the experiments on the KTH dataset that the proposed representation for videos is very suitable in describing the actions and it can give promising performance in human action recognition application.

Table 3 summarizes the overall accuracy of our method in comparison with those reported by Rodriguez et al. [27], Liu et al. [20] and Wang et al. [34]. Wang et al. adopted several kinds of local ST detectors and descriptors [34]. Here, we only choose the same detector and descriptor as ours for comparison. Liu et al. utilized multi-kernel classifier for

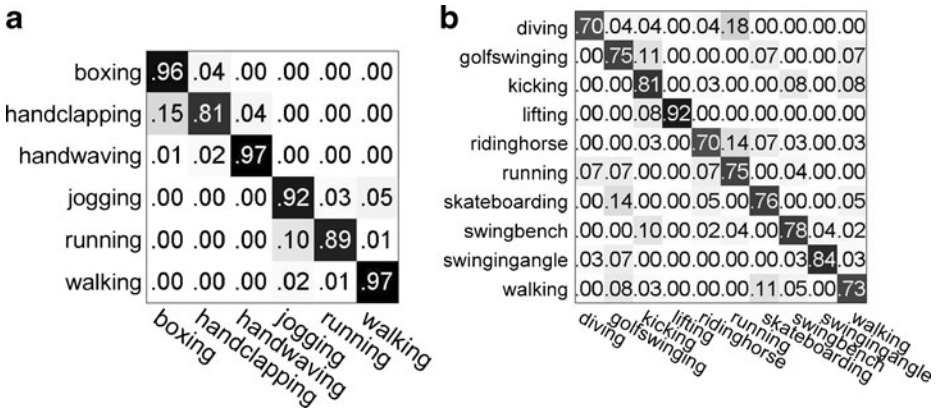

Fig. 10 Classification confusion matrixes on **a** KTH and **b** UCF sports

combining static and motion features [20]. Also, we only compare the performance of motion feature with ours.

As the UCF dataset is comprised of realistic videos containing more noises from moving backgrounds, Liu et al. [20] adopted a feature filtering method by region of interest (ROI) to remove some features in background. We also adopt the ROI selection method of type A by Liu et al. [20] (since we focus on motion features based action recognition in this work). Since the experiment setup is not exactly the same, we cannot compare directly with [20]. However, this experiment demonstrates that our method can deliver good performance with some preprocessing steps for complex realistic videos.

Figure 10 shows the confusion matrixes of the proposed method with JSD on the two datasets. The most confusing category pairs in the KTH dataset include "handclapping" and "boxing", "running" and "jogging". The most confusing category pairs in the UCF sports dataset include "ridinghorse" and "running", "walking" and "skateBoarding", "diving" and "running". It is to be expected because they share similar motion patterns.

5 Conclusion

We explored a localized, continuous and probabilistic video representation for human action recognition. The proposed representation exploited the probabilistic distribution to encode the visual-motion information of an ensemble of local ST features in a continuous and localized manner. Furthermore, based on this probabilistic video representation, the distance of videos was measured in an information-theoretic formulation. This makes the representation compatible with most discriminative classifiers, such as the nearest neighbor schemes and the kernel classifiers. The testing on the KTH and the UCF sports datasets showed that the proposed approach could deliver promising results.

Several issues are worthy of further investigation. First, employing multiple features for video representation in a probabilistic manner could benefit the recognition system, as multiple features provide information redundancy and complementariness. Second, how to encode position information of local features in our representation scheme is our another future research direction.

Acknowledgments This work was supported by National Basic Research Program of China (973 Program, 2007CB311105); National Nature Science Foundation of China (60873165); Co-building Program of Beijing Municipal Education Commission.

References

1. Ballan L, Bertini M, Del Bimbo A, Seidenari L, Serra G (2009) Effective Codebooks for Human Action Categorization. Proceedings of International Conference on Computer Vision 506–513
2. Bishop C (1995) Neural networks for pattern recognition. Oxford University Press, New York
3. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022
4. Bobick AF, Davis JW (2001) The recognition of human movement using temporal templates. IEEE Trans Pattern Anal Mach Intell 23(3):257–267
5. Chan AB, Vasconcelos N, Moreno PJ (2004) A family of probabilistic kernels based on information divergence. Technical Report, University of California, San Diego
6. Do MN, Vetterli M (2002) Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance. IEEE Trans Image Process 11(2):146–158

7. Doll'ar P, Rabaud V, Cottrell G, Belongie S (2005) Behavior recognition via sparse spatio-temporal features. Proceeding of IEEE international workshop on Visual Surveillance Performance Evaluation and Tracking Surveillance 65–72
8. Goldberger J, Gordon S, Greenspan H (2003) An efficient image similarity measure based on approximations of KL-divergence between two Gaussian mixtures. Proceedings of International Conference on Computer Vision 487–493
9. Greenspan H, Goldberger J, Mayer A (2004) Probabilistic space-time video modeling via piecewise GMM. IEEE Trans Pattern Anal Mach Intell 26(3):384–396
10. Greenspan H, Goldberger J, Ridel L (2001) Continuous probabilistic framework for image matching. Comput Vis Image Underst 84(3):384–406
11. Hershey JR, Olsen PA (2007) Approximating the Kullback Leibler divergence between Gaussian mixture models. Proceeding of International Conference on Acoustics, Speech and Signal Processing 4:317–320
12. Hofmann T (1999) Probabilistic latent semantic indexing. In Proceedings of Annual International ACM SIGIR Conference on Research and Development in Information Retrieval 50–57
13. Kendall D (1984) Shape manifolds, procrustean metrics and complex projective spaces. Bull Lond Math Soc 16:81–121
14. Kl'aser A, Marszałek M, Schmid C (2008) A spatio-temporal descriptor based on 3D-gradients. Proceeding of British Machine Vision Conference 995–1004
15. Kullback S (1968) Information theory and statistics. Dover, New York
16. Laptev I, Lindeberg T (2003) Space-time interest points. Proceedings of International Conference on Computer Vision 1:432–439
17. Laptev I, Marszalek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. Proceedings of International Conference on Computer Vision and Pattern Recognition 1–8
18. Cao LL, Liu ZC, Huang TS (2010) Cross-dataset action detection. Proceeding of International Conference on Computer Vision and Pattern Recognition 1998–2005
19. Lin J (1991) Divergence measures based on the Shannon entropy. IEEE Trans Inf Theory 37:145–151
20. Liu JG, Luo JB, Shah M (2009) Action recognition in unconstrained amateur videos. Proceeding of International Conference on Acoustics, Speech and Signal Processing 3549–3552
21. Liu JG, Shah M (2008) Learning human actions via information maximization. Proceeding of International Conference on Computer Vision and Pattern Recognition 1–8
22. Liu JG, Yang Y, Shah M (2009) Learning semantic visual vocabularies using diffusion distance. Proceedings of International Conference on Computer Vision and Pattern Recognition 461–468
23. Lowe D (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vision 60(2):91–110
24. Niebles JC, Wang HC, Li FF (2008) Unsupervised learning of human action categories using spatial-temporal words. Int J Comput Vision 79:299–318
25. Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77:257–286
26. Rissanen J (1978) Modeling by shortest data description. Automatic 14:465–471
27. Rodriguez MD, Ahmed J, Shah M (2008) Action MACH: a spatio-temporal maximum average correlation height filter for action recognition. Proceeding of International Conference on Computer Vision and Pattern Recognition 1–8
28. Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local SVM approach. Proceeding of International Conference on Pattern Recognition 3:32–36
29. Scovanner P, Ali S, Shah M (2007) A 3-dimensional SIFT descriptor and its application to action recognition. In ACM International Conference on Multimedia 357–360
30. Song Y, Tang S, Zheng YT, Chua TS, Zhang YD, Lin SX (2010) A distribution based video representation for human action recognition. In Proceedings of IEEE International Conference on Multimedia & Expo
31. Vasconcelos N, Ho P, Moreno P (2004) The Kullback-Leibler kernel as a framework for discriminant and localized representation for visual recognition. Proceedings of European Conference on Computer Vision 430–441
32. Veeraraghavan A, Roy-Chowdhury AK, Chellappa R (2005) Matching shape sequences in video with applications in human movement analysis. IEEE Trans Pattern Anal Mach Intell 27(12):1896–1909
33. Vergés-Llahí J, Sanfeliu (2005) A evaluation of distances between color image segmentations. Pattern Recognit Image Anal 263–270
34. Wang H, Uiiah MM, Klaser A, Laptev I, Schmid C (2009) Evaluation of local spatio-temporal features for action recognition. Proceeding of British Machine Vision Conference 127–138
35. Willems G, Tuytelaars T, Van Gool L (2008) An efficient dense and scale-invariant spatio-temporal interest point detector. Proceedings of European Conference on Computer Vision 650–663

36. Wong S-F, Cipolla R (2007) Extracting spatiotemporal interest points using global information. Proceedings of International Conference on Computer Vision 1–8
37. Xiong ZY, Radhakrishnan R, Divakaran A, Huang TS (2004) Effective and efficient sports highlights extraction using the minimum description length criterion in selecting GMM structures. Proceeding of International Conference on Multimedia and Expo 3:1947–1950
38. Xu LM, Tang ZM (2007) Speaker identification using multi-step clustering algorithm with transformation based GMM. Autom Control Comput Sci 41(4):224–231
39. Zhou X, Zhuang XD, Yan SC, Chang SF, Johnson MH, Huang TS (2008) SIFT-Bag kernel for video event analysis. In ACM International Conference on Multimedia 229–238

Yan Song born in 1983, is currently working toward her Ph.D. degree in the Multimedia Computing Group, Advanced Computing Research Lab, Institute of Computing Technology, Chinese Academy of Sciences. She received the B.S. degree in Nanjing University of Science and Technology, Nanjing, Jiangsu, China, in 2005. Her research interests focus on multimedia information processing, in particular, on video content analysis and understanding. The main focus is on the human action recognition and event detection of video.

Dr. Sheng Tang is an associate professor in the Institute of Computing Technology, Chinese Academy of Sciences (ICT-CAS), China. He received his Ph.D. degree in computer application technology at ICT-CAS in 2006. He visited National University of Singapore (NUS) for participating in TREC Video Retrieval Evaluation (TRECVID) tasks from July to August in 2006. From 2006 to 2008, he was TRECVID team

(MCG-ICT-CAS) leader of the Multimedia Computing Group at ICT-CAS. From Feb., 2009 to Feb., 2010, he worked as a visiting research fellow in NUS under the instruction of Prof. Chua Tat-Seng. His current research interests are in the fields of pattern recognition and content-based multimedia retrieval and indexing. He served as the reviewer for Journal of Visual Communication and Image Representation, Multimedia Tools and Applications, Journal of Computer Science and Technology

Dr. Yan-Tao Zheng is a research engineer at Institute for Infocomm Research (I2R), Singapore. He received his Ph.D from National University of Singapore and B.Eng (with 1st class Hons) from Nanyang Technological University, Singapore. His research interests include geo-mining in multimedia, image annotation and video search. He is the recipient of a number of international awards, including Champion of Star Challenge, Microsoft Research Fellowship, IBM Waston Emerging Multimedia Leaders, and so on. During his attachment at Google Inc. in 2008, he developed a world-scale landmark recognition engine together with Google engineers, which has been highly praised and well publicized. He has served as program committee member and reviewer of a number of prestigious international conferences and journals.

Dr. Chua Tat-Seng is the Professor at the School of Computing, National University of Singapore. He was the Acting and Founding Dean of the School of Computing from 1998–2000. He spent three years as a research staff member at the Institute of Systems Science (now I2R) in late 1980s. His main research interest is in multimedia information processing, in particular, on the extraction, retrieval and question-answering (QA) of video and text information. He focuses on the use of relations between entities and external

information & knowledge sources to enhance information processing. His current projects include: news video retrieval and tracking, question answering (QA), video QA, and information extraction on the web. His group participates regularly in TREC-QA and TRECVID news video retrieval evaluations. He obtained his PhD from the University of Leeds, UK.

Dr. Yongdong Zhang born in 1973, received his Ph.D. degree in Electronic Engineering from Tianjin University, Tianjin, China, in 2002. He is an Associate Professor and Professor (in 2009) with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His research interests are in the field of video coding and transcoding, video analysis and retrieval, and universal media access.

Dr. Shouxun Lin born in 1948, received his Ph.D. degree from Beijing University of Technology, Beijing, China, in 1998. Since 1990, he has been Associate Professor and Professor (in 1995) with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His research interests include multimedia processing and comparison, video coding, video analysis, multimedia indexing, statistical machine translation, and evaluation of computer human interaction.