Motion Context: A New Representation for Human Action Recognition

Ziming Zhang, Yiqun Hu, Syin Chan, and Liang-Tien Chia

Center for Multimedia and Network Technology, School of Computer Engineering,
Nanyang Technological University, Singapore 639798
{zhan0154,yqhu,asschan,asltchia}@ntu.edu.sg

Abstract. One of the key challenges in human action recognition from video sequences is how to model an action sufficiently. Therefore, in this paper we propose a novel motion-based representation called ***Motion Context*** (MC), which is insensitive to the scale and direction of an action, by employing image representation techniques. A MC captures the distribution of the *motion words* (MWs) over relative locations in a local region of the *motion image* (MI) around a reference point and thus summarizes the local motion information in a rich 3D MC descriptor. In this way, any human action can be represented as a 3D descriptor by summing up all the MC descriptors of this action. For action recognition, we propose 4 different recognition configurations: MW+pLSA, MW+SVM, MC+w^3-pLSA (a new direct graphical model by extending pLSA), and MC+SVM. We test our approach on two human action video datasets from KTH and Weizmann Institute of Science (WIS) and our performances are quite promising. For the KTH dataset, the proposed MC representation achieves the highest performance using the proposed w^3-pLSA. For the WIS dataset, the best performance of the proposed MC is comparable to the state of the art.

1 Introduction

With the development of advanced security systems, human action recognition in video sequences has become an important research topic in computer vision, whose aim is to make machines recognize human actions using different types of information, especially the motion information, in the video sequences.

The basic process for this problem can be divided into three issues: First, how to detect the existence of human actions? Second, how to represent human actions? Lastly, how to recognize these actions? Many research works have been done to address these issues (e.g. [1], [2], [3], [4], [5], [6]). In this paper, we mainly focus on the second issue, that is, how to represent human actions after having detected their existence. In our approach, we model each video sequence as a collection of so-called *motion images* (MIs), and to model the action in each MI, we propose a novel motion-based representation called ***motion context*** (MC), which is insensitive to the scale and direction of an action, to capture the distribution of the *motion words* (MWs) over relative locations in a local

D. Forsyth, P. Torr, and A. Zisserman (Eds.): ECCV 2008, Part IV, LNCS 5305, pp. 817–829, 2008.

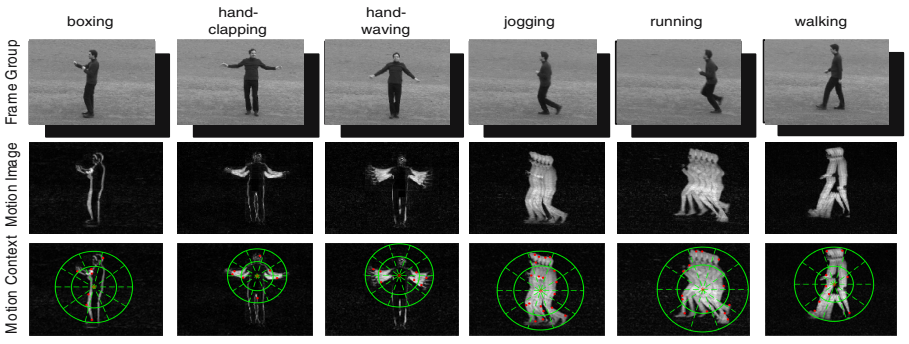

Fig. 1. Illustrations of the frame groups, motion images, and our motion context representations on the KTH dataset. This figure is best viewed in color.

region around a reference point and thus summarize the local motion information in a rich, local 3D MC descriptor. Fig.1 illustrates some MIs and their corresponding MC representations using the video clips in the KTH dataset. To describe an action, only one 3D descriptor is generated by summing up all the MC descriptors of this action in the MIs. For action recognition, we employ 3 different approaches: pLSA [7], w^3-pLSA (a new direct graphical model by extending pLSA) and SVM [8]. Our approach is tested on two human action video datasets from KTH [2] and Weizmann Institute of Science [9], and the performances are quite promising.

The rest of this paper is organized as follows: Section 2 reviews some related works in human action recognition. Section 3 presents the details of our MC representation. Section 4 introduces the 3 recognition approaches. Our experimental results are shown in Section 5, and finally Section 6 concludes the paper.

2 Related Work

Each video sequence can be considered as a collection of consecutive images (frames), which makes it possible to model human actions using some image representation techniques. One influential model is the *Bag-of-Words* (BOW) model (e.g. [4], [6], [10], [11]). This model represents each human action as a collection of independent codewords in a pre-defined codebook generated from the training data. However, videos contain temporal information while images do not. So how to exploit this temporal information becomes a key issue for human action representation.

Based on image representation techniques, many research works have shown that temporal information can be integrated with the interesting point detectors and descriptors to locate and describe the interesting points in the videos. Laptev et al. [1] proposed a 3D interesting point detector where they added the temporal constraint to the Harris interesting point detector to detect local structures in the space-time dimensions. Efros et al. [12] proposed a motion descriptor using

the optical flow from different frames to represent human actions. Recently, Scovanner et al. [4] applied sub-histograms to encode local temporal and spatial information to generate a 3D version of SIFT [13] (3D SIFT), and Savarese et al. [14] proposed so-called "spatial-temporal correlograms" to encode flexible long range temporal information into the spatial-temporal motion features.

However, a common issue behind these interesting point detectors is that the detected points sometimes are too few to sufficiently characterize the human action behavior, and hence reduce the recognition performance. This issue has been avoided in [6] by employing the separable linear filter method [3], rather than such space-time interesting point detectors, to obtain the motion features using a quadrature pair of 1D Gabor filters temporally.

Another way of using temporal information is to divide a video into smaller groups of consecutive frames as the basic units and represent a human action as a collection of the features extracted from these units. In [15], [5], every three consecutive frames in each video were grouped together and integrated with their graphical models as a node to learn the spatial-temporal relations among these nodes. Also in [16], the authors took the average of a sequence of binary silhouette images of a human action to create the "Average Motion Energy" representation. Similarly, [17] proposed a concept of "Motion History Volumes", an extension of "Motion History Images" [18], to capture the motion information from a sequence of video frames.

After the human action representations have been generated, both discriminative approaches (e.g. kernel approaches [2]) and generative approaches (e.g. pLSA [19], MRF [15], [5], semi-LDA [10], hierarchical graphical models [6]) can be employed to recognize them.

3 Motion Context Representation

A *motion context* representation is generated based on the *motion words* which are extracted from the *motion images*.

3.1 Motion Image

We believe that effective utilization of the temporal information is crucial for human action recognition. In our approach, we adopt the strategy in [17], that is, to group the consecutive frames of each video sequence according to their temporal information.

More specifically, to generate a ***motion image*** (MI), first $U * V$ frames of a video sequence are extracted, converted into gray scale and divided into non-overlapping U groups, each with V consecutive frames. Then we calculate the standard deviation (stdev) among the frames within a group pixel by pixel to detect the motion information. Finally, putting the stdev values into the corresponding pixel positions, a MI is generated for each frame group. Fig.2 illustrates the MI generation process for a frame group. Motions usually cause strong changes in the pixel intensity values at the corresponding positions among the

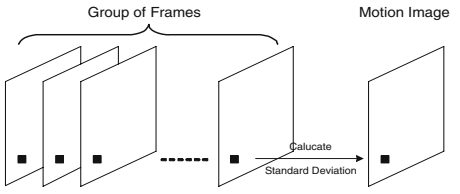

Fig. 2. Illustration of the MI generation process for a frame group. The black dots denote the pixel intensity values.

consecutive frames. Since stdev can measure the variances of the pixel intensity values, it can definitely detect motions.

We would like to mention that the length of each group, V, should be long enough to capture the motion information sufficiently but not too long. Fig.3 illustrates the effects of different V on the MIs of human running and walking. If $V = 5$, the difference between the two actions is quite clear. With V increased to 60, the motion information of both actions spreads in the MIs, making it difficult to distinguish them. A further investigation of V will be essential in our MC representation.

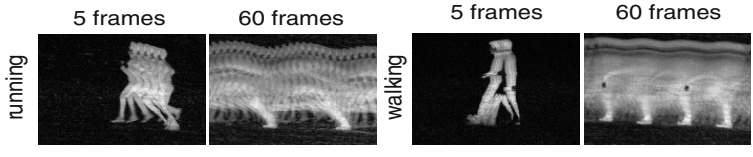

Fig. 3. Illustration of effects of different lengths of frame groups on the MIs using human running and walking

3.2 Motion Word

The concept of ***motion words*** (MWs) refers to that of visual words in the BOW model. After generating the MIs, some image interesting point detectors are first applied to locate the important patches in the MIs. Then image descriptors are employed to map these patches into a high dimensional feature space to generate local feature vectors for them. Next, using clustering approaches such as K-means, these local feature vectors in the training data are clustered to generate a so-called *motion word dictionary* where the centers of the clusters are treated as the MWs.

3.3 Motion Context

For each MW, there is one important affiliated attribute, its location in the corresponding MI. For human action recognition, the relative movements of different parts of the body are quite useful. To capture the structures of these relative movements, we introduce the concept of ***motion context*** (MC). This concept

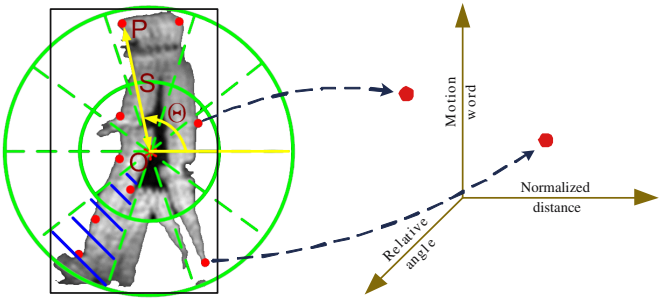

Fig. 4. Illustration of our MC representation (left) and its 3D descriptor (right). On the left, P denotes a MW at an interesting point, O denotes the reference point, Θ and S denote the relative angle and normalized distance between P and O in the support region (the black rectangle), respectively, and the shaded sector (blue) denotes the orientation of the whole representation. On the right, each MW is quantized into a point to generate a 3D MC descriptor. This figure is best viewed in color.

is inspired by Shape Context (SC) [20], which has been widely used in object recognition. The basic idea of SC is to locate the distribution of other shape points over relative positions in a region around a pre-defined reference point. Subsequently, 1D descriptors are generated to represent the shapes of objects.

In our representation, we utilize the polar coordinate system to capture the relative angles and distances between the MWs and the reference point (the pole of the polar coordinate system) for each action in the MIs, similar to SC. This reference point is defined as the geometric center of the human motion, and the relative distances are normalized by the maximum distance in the support region, which makes the MC insensitive to changes in scale of the action. Here, the support region is defined as the area which covers the human action in the MI. Fig.4 (left) illustrates our MC representation. Suppose that the angular coordinate is divided into M equal bins, the radial coordinate is divided into N equal bins and there are K MWs in the dictionary, then each MW can be put into one of the $M*N$ bins to generate a 3D MC descriptor for each MC representation, as illustrated in Fig.4 (right). To represent a human action in each video sequence, we sum up all the MC descriptors of this action to generate one 3D descriptor with the same dimensions.

When generating MC representations, another factor should also be considered, that is, the direction of the action, because the same action may occur in different directions. E.g. a person may be running in one direction or the opposite direction. In such cases, the distributions of the interesting points in the two corresponding MIs should be roughly symmetric about the y-axis. Combining the two distributions for the same action will reduce the discriminability of our representation. To avoid this, we define the orientation of each MC representation as the sector where most interesting points are detected, e.g. the shaded one (blue) in Fig. 4 (left). This sector can be considered to represent the main characteristics of the motion in one direction. For the same action but in the

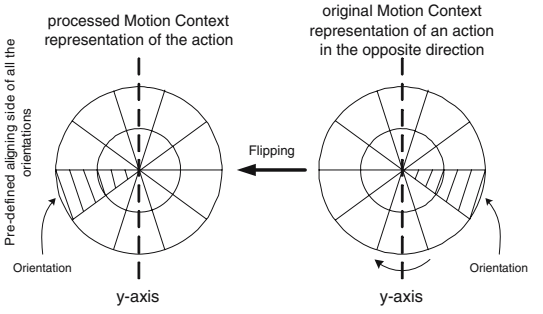

Fig. 5. Illustration of aligning an inconsistent MC representation of an action in the opposite direction. The pre-defined orientation of the actions is the left side of y-axis.

opposite direction, we then align all the orientations to the pre-defined side by flipping the MC representations horizontally around the y-axis. Thus our representation is symmetry-invariant. Fig.5 illustrates this process. Notice that this process is done automatically without the need to know the action direction.

The entire process of modeling human actions using the MC representation is summarized in Table 1.

Table 1. The main steps of modeling the human actions using the MC representation

Step 1 Obtain the MIs from the video sequences.
Step 2 Generate the MC representation for each human action in the MIs.
Step 3 Generate the 3D MC descriptor for each MC representation.
Step 4 Sum up all the 3D MC descriptors of an action to generate one 3D descriptor to represent this action.

4 Action Recognition Approaches

We apply 3 different approaches to recognize the human actions based on the MWs or the 3D MC descriptors: pLSA, w^3-pLSA and SVM.

4.1 pLSA

pLSA aims to introduce an aspect model, which builds an association between documents and words through the latent aspects by probability. Here, we follow the terminology of text classification where pLSA was used first. The graphical model of pLSA is illustrated in Fig.6 (a).

Suppose $D = \{d_1, \ldots, d_I\}$, $W = \{w_1, \ldots, w_J\}$ and $Z = \{z_1, \ldots, z_K\}$ denote a document set, a word set and a latent topic set, respectively. pLSA models the joint probability of documents and words as:

$$P(d_i, w_j) = \sum_k P(d_i, w_j, z_k) = \sum_k P(w_j|z_k)P(z_k|d_i)P(d_i) \tag{1}$$

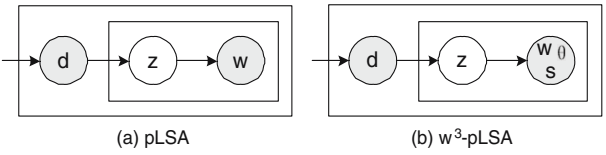

Fig. 6. Graphical models of pLSA (a) and our w^3-pLSA (b)

where $P(d_i, w_j, z_k)$ denotes the joint probability of document d_i, topic z_k and word w_j, $P(w_j|z_k)$ denotes the probability of w_j occurring in z_k, $P(z_k|d_i)$ denotes the probability of d_i classified into z_k, and $P(d_i)$ denotes the prior probability of d_i modeled as a multinomial distribution.

Furthermore, pLSA tries to maximize the $\mathcal{L}$ function below:

$$\mathcal{L} = \sum_i \sum_j n(d_i, w_j) \log P(d_i, w_j) \tag{2}$$

where $n(d_i, w_j)$ denotes the document-word co-occurrence table, where the number of co-occurrences of d_i and w_j is recorded in each cell.

To learn the probability distributions involved, pLSA employs the Expectation Maximization (EM) algorithm shown in Table 2 and records $P(w_j|z_k)$ for recognition, which is learned from the training data.

Table 2. The EM algorithm for pLSA

E-step:
$P(z_k\|d_i, w_j) \propto P(w_j\|z_k)P(z_k\|d_i)P(d_i)$
M-step:
$P(w_j\|z_k) \propto \sum_i n(d_i, w_j)P(z_k\|d_i, w_j)$
$P(z_k\|d_i) \propto \sum_j n(d_i, w_j)P(z_k\|d_i, w_j)$
$P(d_i) \propto \sum_j n(d_i, w_j)$

4.2 w^3-pLSA

To bridge the gap between the human actions and our MC descriptors, we extend pLSA to develop a new graphical model, called w^3-*pLSA*. See Fig.6 (b), where $\boldsymbol{d}$ denotes human actions, $\boldsymbol{z}$ denotes latent topics, $\boldsymbol{w}$, θ and $\boldsymbol{s}$ denote motion words, and the indexes in the angular and radial coordinates in the polar coordinate system, respectively.

Referring to pLSA, we model the joint probability of human actions, motion words and their corresponding indices in the angular and radial coordinates as

$$P(d_i, w_j, \theta_m, s_r) = \sum_k P(d_i, w_j, \theta_m, s_r, z_k) = \sum_k P(d_i)P(z_k|d_i)P(w_j, \theta_m, s_r|z_k) \tag{3}$$

and maximize the $\widehat{\mathcal{L}}$ function below.

$$\widehat{\mathcal{L}} = \sum_i \sum_j \sum_m \sum_r n(d_i, w_j, \theta_m, s_r) \log P(d_i, w_j, \theta_m, s_r) \tag{4}$$

Similarly, to learn the probability distributions involved, w^3-pLSA employs the Expectation Maximization (EM) algorithm shown in Table 3 and records $P(w_j, \theta_m, s_r|z_k)$ for recognition, which is learned from the training data.

Table 3. The EM algorithm for w^3-pLSA

E-step:
$P(z_k\|d_i, w_j, \theta_m, s_r) \propto P(w_j, \theta_m, s_r\|z_k)P(z_k\|d_i)P(d_i)$
M-step:
$P(w_j, \theta_m, s_r\|z_k) \propto \sum_i n(d_i, w_j, \theta_m, s_r)P(z_k\|d_i, w_j, \theta_m, s_r)$
$P(z_k\|d_i) \propto \sum_{j,m,r} n(d_i, w_j, \theta_m, s_r)P(z_k\|d_i, w_j, \theta_m, s_r)$
$P(d_i) \propto \sum_{j,m,r} n(d_i, w_j, \theta_m, s_r)$

4.3 Support Vector Machine

A support vector machine (SVM) [8] is a powerful tool for binary classification tasks. First it maps the input vectors into a higher dimensional feature space, then it conducts a separating hyperplane to separate the input data, finally on each side of this hyperplane two parallel hyperplanes are conducted. SVM tries to find the separating hyperplane which maximizes the distance between the two parallel hyperplanes. Notice that in a SVM, there is an assumption that the larger the distance between the two parallel hyperplanes the smaller the generalization error of the classifier will be.

Specifically, suppose the input data is $\{(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), \cdots, (\mathbf{x_n}, y_n)\}$ where $\mathbf{x_i}(i = 1, 2, \cdots, n)$ denotes the input vector and the corresponding $y_i(i = 1, 2, \cdots, n)$ denotes the class label (positive "1" and negative "-1"). Then the separating hyperplane is defined as $\mathbf{w} \cdot \mathbf{x} + b = 0$ and the two corresponding parallel hyperplanes are $\mathbf{w} \cdot \mathbf{x} + b = 1$ for the positive class and $\mathbf{w} \cdot \mathbf{x} + b = -1$ for the negative class, where $\mathbf{w}$ is the vector perpendicular to the separating hyperplane and b is a scalar. If a test vector $\mathbf{x_t}$ satisfies $\mathbf{w} \cdot \mathbf{x_t} + b > 0$, it will be classified as a positive instance. Otherwise, if it satisfies $\mathbf{w} \cdot \mathbf{x_t} + b < 0$, it will be classified as a negative instance. A SVM tries to find the optimal $\mathbf{w}$ and b to maximize the distance between the two parallel hyperplanes.

5 Experiments

Our approach has been tested on two human action video datasets from KTH [2] and Weizmann Institute of Science (WIS) [9]. The KTH dataset is one of the largest datasets for human action recognition containing six types of human actions: boxing, handclapping, handwaving, jogging, running, and walking. For

each type, there are 99 or 100 video sequences of 25 different persons in 4 different scenarios: outdoors (S1), outdoors with scale variation (S2), outdoors with different clothes (S3) and indoors (S4), as illustrated in Fig.7 (left). In the WIS dataset, there are altogether 10 types of human actions: walk, run, jump, gallop sideways, bend, one-hand wave, two-hands wave, jump in place, jumping jack, and skip. For each type, there are 9 or 10 video sequences of 9 different persons with the similar background, as shown in Fig.7 (right).

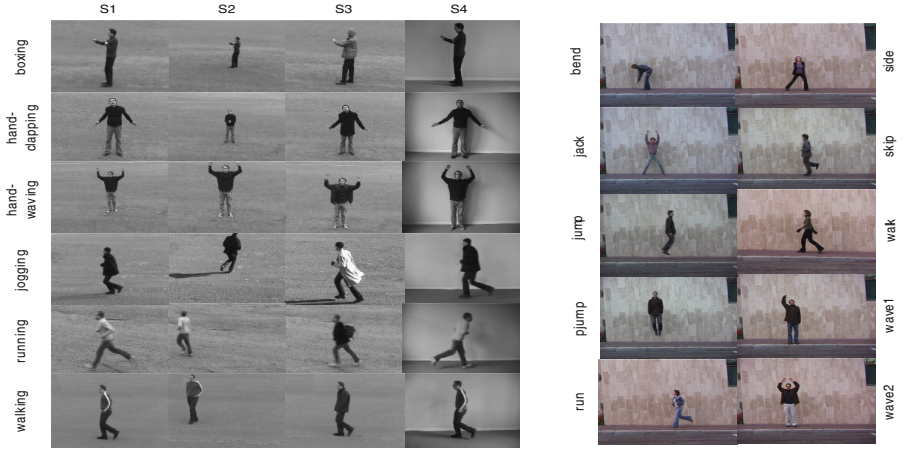

Fig. 7. Some sample frames from the KTH dataset (left) and the WIS dataset (right)

5.1 Implementation

To generate MC representations for human actions, we need to locate the reference points and the support regions first. Some techniques in body tracking (e.g. [21]) can be applied to locate the areas and the geometric centers of the human bodies in each frame group of a video sequence. The integration of the areas of a person can be defined as its support region and the mean of its centers can be defined as the reference point for this action in the MI. However, this issue is beyond the purpose of this paper. So considering that in our datasets each video sequence only contains one person, we simply assume that in each MI the support region of each human action covers the whole MI, and we adopted a simple method to roughly locate the reference points. First, we generated one MI from every 5-frame group of each video sequence empirically. Then a Gaussian filter was applied to denoise these MIs so that the motion information from the background was suppressed. Next, we used the Canny edge detector to locate the edges in each MI, and finally took the geometric center of the edge points as the reference point for the action.

After locating the reference points, we followed the steps in Table 1 to generate the MC representations for human actions. The detector and descriptor involved in Step 2 are the Harris-Hessian-Laplace detector [22] and the SIFT descriptor

Table 4. Comparison (%) between our approach and others on the KTH dataset

Rec.Con.	Tra.Str.	boxing	hand-c	hand-w	jogging	running	walking	average
MW+pLSA	SDE	85.2	91.9	91.7	71.2	73.6	82.1	82.62
	LOO	82.0	90.9	91.0	82.0	79.0	83.0	84.65
MW+SVM	SDE	90.4	84.8	82.8	65.1	76.1	82.0	80.20
	LOO	85.0	82.8	82.0	62.0	70.0	87.0	78.14
MC+w^3-pLSA	SDE	98.4	90.8	93.9	79.3	77.9	91.7	88.67
	LOO	95.0	97.0	93.0	88.0	84.0	91.0	***91.33***
MC+SVM	SDE	91.7	91.6	88.1	78.0	84.7	90.4	87.42
	LOO	88.0	93.9	91.0	77.0	85.0	90.0	87.49
Savarese et al. [14]	LOO	97.0	91.0	93.0	64.0	83.0	93.0	86.83
Wang et al. [10]	LOO	96.0	97.0	100.0	54.0	64.0	99.0	85.00
Niebles et al. [19]	LOO	100.0	77.0	93.0	52.0	88.0	79.0	81.50
Dollár et al. [3]	LOO	93.0	77.0	85.0	57.0	85.0	90.0	81.17
Schuldt et al. [2]	SDE	97.9	59.7	73.6	60.4	54.9	83.8	71.72
Ke et al. [24]	SDE	69.4	55.6	91.7	36.1	44.4	80.6	62.96
Wong et al. [25]	SDE	96.0	92.0	83.0	79.0	54.0	100.0	84.00

[13], and the clustering method used here is K-means. Then based on the MWs and the MC descriptors of the training data, we trained pLSA, w^3-pLSA and SVM for each type of actions separately, and a test video sequence was classified to the type of actions with the maximum likelihood.

5.2 Experimental Results

To show the efficiency of our MC representation and the discriminability of the MWs, we designed 4 different recognition configurations: MW+pLSA, MW+SVM, MC+w^3-pLSA, and MC+SVM. Here we used libsvm [23] with the linear kernel. To utilize the MWs, we employed the BOW model to represent each human action as a histogram of the MWs without the $M*N$ spatial bins.

First, we tested our approach on the KTH dataset. We adopted two different training strategies: split-data-equally (SDE) and leave-one-out (LOO). The SDE strategy means that the video collection is divided into two equal sets randomly: one as the training data (50 video sequences) and the other as the test data for each type of actions, and we repeated this experiment for 15 times. In the LOO strategy, for each type of actions, only the video sequences of one person are selected as the test data and the rest as the training data, and when applying this strategy to the KTH dataset, for each run we randomly selected one person for each type of actions as the test data and repeated this experiment for 15 times. Empirically, in our model, the number of MWs is 100, and the numbers of the quantization bins in the angular and radial dimensions are 10 and 2, respectively. The number of latent topics in both graphical models is 40.

Table 4 shows our average recognition rate for each type of actions and the comparison with others on the KTH dataset under different training strategies and recognition configurations. From this table, we can draw the following

Table 5. Comparison (%) between our approach and others on the WIS dataset. Notice that "×" denotes that this type of actions was not involved in their experiments.

Rec.Con.	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2	ave.
MW+pLSA	77.8	100.0	88.9	88.9	70.0	100.0	60.0	100.0	66.7	88.9	84.1
MW+SVM	100.0	100.0	100.0	77.8	30.0	77.8	40.0	100.0	100.0	100.0	81.44
MC+w^3-pLSA	66.7	100.0	77.8	66.7	80.0	88.9	100.0	100.0	100.0	100.0	88.0
MC+SVM	100.0	100.0	100.0	88.9	80.0	100.0	80.0	80.0	100.0	100.0	***92.89***
Wang et al. [16]	100.0	100.0	89.0	100.0	100.0	100.0	89.0	100.0	89.0	100.0	96.7
Ali et al. [26]	100.0	100.0	55.6	100.0	88.9	88.9	×	100.0	100.0	100.0	92.6
Scovanner [4]	100.0	100.0	67.0	100.0	80.0	100.0	50.0	89.0	78.0	78.0	84.2
Niebles et al. [6]	100.0	100.0	100.0	44.0	67.0	78.0	×	56.0	56.0	56.0	72.8

conclusions: (1) MWs without any spatial information are not discriminative enough to recognize the actions. MW+pLSA returns the best performance (84.65%) using MWs, which is lower than the state of the art. (2) MC representation usually achieves better performances than MWs, which demonstrates that the distributions of the MWs are quite important for action recognition. MC+w^3-pLSA returns the best performance (91.33%) among all the approaches.

Unlike the KTH dataset, the WIS dataset only has 9 or 10 videos for each type of human actions, which may result in underfit when training the graphical models. To utilize this dataset sufficiently, we only used the LOO training strategy to learn the models for human actions and tested on all the video sequences. We compare our average recognition rates with others in Table 5. The experimental configuration of the MC representation is kept the same as that used on the KTH dataset, while the number of MWs used in the BOW model is modified empirically to 300. The number of latent topics is unchanged. From this table, we can see that MC+SVM still returns the best performance (92.89%) among the different configurations, which is comparable to other approaches and higher than the best performance (84.1%) using MW. These results demonstrate that our MC presentation can model the human actions properly with the distributions of the MWs.

6 Conclusion

We have demonstrated that our ***Motion Context*** (MC) representation, which is insensitive to changes in the scales and directions of the human actions, can model the human actions in the *motion images* (MIs) effectively by capturing the distribution of the *motion words* (MWs) over relative locations in a local region around the reference point and thus summarize the local motion information in a rich 3D descriptor. To evaluate this novel representation, we adopt two training strategies (split-data-equally (SDE) and leave-one-out (LOO)), design 4 different recognition configurations (MW+pLSA, MW+SVM, MC+w^3-pLSA, and MC+SVM) and test them on two human action video datasets from KTH

and Weizmann Institute of Science (WIS). The performances are promising. For the KTH dataset, all configurations using MC outperform existing approaches where the best performances are obtained using w^3-pLSA (88.67% for SDE and 91.33% for LOO). For the WIS dataset, our MC+SVM returns the comparable performance (92.89%) using the LOO strategy.

References

1. Laptev, I., Lindeberg, T.: Space-time interest points. In: ICCV (2003)
2. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: ICPR 2004, vol. III, pp. 32–36 (2004)
3. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS (October 2005)
4. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. ACM Multimedia, 357–360 (2007)
5. Wang, Y., Loe, K.F., Tan, T.L., Wu, J.K.: Spatiotemporal video segmentation based on graphical models. Trans. IP 14, 937–947 (2005)
6. Niebles, J., Fei Fei, L.: A hierarchical model of shape and appearance for human action classification. In: CVPR 2007, pp. 1–8 (2007)
7. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. In: Mach. Learn., Hingham, MA, USA, vol. 42, pp. 177–196. Kluwer Academic Publishers, Dordrecht (2001)
8. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. In: Data Mining and Knowledge Discovery, vol. 2, pp. 121–167 (1998)
9. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: ICCV 2005, vol. II, pp. 1395–1402 (2005)
10. Wang, Y., Sabzmeydani, P., Mori, G.: Semi-latent dirichlet allocation: A hierarchical model for human action recognition. In: HUMO 2007, pp. 240–254 (2007)
11. Ikizler, N., Duygulu, P.: Human action recognition using distribution of oriented rectangular patches. In: HUMO 2007, pp. 271–284 (2007)
12. Efros, A., Berg, A., Mori, G., Malik, J.: Recognizing action at a distance. In: ICCV 2003, pp. 726–733 (2003)
13. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 20, 91–110 (2003)
14. Savarese, S., Sel Pozo, A., Fei-Fei, J.N.L.: Spatial-temporal correlations for unsupervised action classification. In: IEEE Workshop on Motion and Video Computing, Copper Mountain, Colorado (2008)
15. Wang, Y., Tan, T., Loe, K.: Video segmentation based on graphical models. In: CVPR 2003, vol. II, pp. 335–342 (2003)
16. Wang, L., Suter, D.: Informative shape representations for human action recognition. In: ICPR 2006, vol. II, pp. 1266–1269 (2006)
17. Weinland, D., Ronfard, R., Boyer, E.: Free viewpoint action recognition using motion history volumes. Computer Vision and Image Understanding 104 (November/December 2006)
18. Bobick, A., Davis, J.: The recognition of human movement using temporal templates. PAMI 23(3), 257–267 (2001)
19. Niebles, J., Wang, H., Wang, H., Fei Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. In: BMVC 2006, vol. III, p. 1249 (2006)

20. Belongie, S., Malik, J., Puzicha, J.: Shape context: A new descriptor for shape matching and object recognition. In: NIPS, pp. 831–837 (2000)
21. Bissacco, A., Yang, M.H., Soatto, S.: Fast human pose estimation using appearance and motion via multi-dimensional boosting regression. In: CVPR (2007)
22. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis & Machine Intelligence 27, 1615–1630 (2005)
23. Chang, C., Lin, C.: Libsvm: a library for support vector machines, Online (2001)
24. Ke, Y., Sukthankar, R., Hebert, M.: Efficient visual event detection using volumetric features. In: International Conference on Computer Vision, vol. 1, p. 166 (October 2005)
25. Wong, S., Kim, T., Cipolla, R.: Learning motion categories using both semantic and structural information. In: CVPR 2007, pp. 1–6 (2007)
26. Ali, S., Basharat, A., Shah, M.: Chaotic invariants for human action recognition. In: ICCV 2007, pp. 1–8 (2007)