Multimed Tools Appl (2016) 75:11847–11864
DOI 10.1007/s11042-015-2661-y

Multi-view transition HMMs based view-invariant human action recognition method

Xiaofei Ji[1] · Zhaojie Ju[2] · Ce Wang[1] · Changhui Wang[1]

Received: 8 January 2015 / Revised: 7 April 2015 / Accepted: 24 April 2015 /
Published online: 10 May 2015

Abstract View-invariant human action recognition is a challenging research topic in computer vision. Hidden Markov Models(HMM) and their extensions have been widely used for view-invariant action recognition. However those methods are usually according to a large parameter space, requiring amounts of training data and with low classification accuracies for real application. A novel graphical structure based on HMM with multi-view transition is proposed to model the human action with viewpoint changing. The model consists of multiple sub action models, which correspond to the traditional HMM utilized to model the human action in a particular rotation viewpoint space. In the training process, the novel model can be built by connecting the sub action models between adjacent viewpoint spaces. In the recognition process, action with unknown viewpoint is recognized by using improved forward algorithm. The proposed model can not only simplify the model training process by decomposing the parameter space into multiple sub-spaces, but also improve the performance the algorithm by constraining the possible viewpoint changing. Experiment results on IXMAS dataset demonstrated that the proposed model obtains better performance than other recent view-invariant action recognition method.

Keywords Action recognition · View-invariant · Multi-view transition · Hidden Markov Model

✉ Xiaofei Ji
jixiaofei7804@126.com

Zhaojie Ju
zhaojie.ju@port.ac.uk

[1] School of Automation, Shenyang Aerospace University, Shenyang, China

[2] School of Computing, University of Portsmouth, Portsmouth, UK

1 Introduction

Visual-based human action recognition has received considerable attention in computer vision during past few years. The growing interest is due to the increasing number of real-world applications such as visual surveillance, human-computer interaction, video indexing and retrieval [9, 17]. However, the data with viewpoint changes are very common and basically inevitable in those real-world scenarios, it is desired that the recognition algorithm exhibit view-invariance [4, 5]. The object of the view-invariant human action recognition is to recognize different actions performed by different actors under different camera viewpoints, with different style regardless of large variation in manner and speed. It also suffers from various factors such as clustered background, occlusion, camera movement and illumination change. So it remains challenging to recognize actions from different viewpoints.

Recent research on view-invariant human action recognition can be characterized by two classes of methods: **template matching methods** and **state-space approaches**. Template based approached focus on extracting low-level image features which are then compared it to the prestored action prototypes during recognition. The view constraint is mostly removed during feature extraction in these recognition methods. A kind of template-based methods directly choose the view-independent motion features [21, 23, 25, 26]. For example, Weinland et al. [21] extracted the view-invariant features from Fourier space of Motion History Volumes (MHVs) for action recognition. The advantages of this kind of method are low computational cost and simple implementation. However, they are usually more sensitive to noise and variance of the time interval of the movements. Another kind of template-based methods estimates the parameters of camera viewpoint according to the human body movement direction. Then the observation information from every frame is projected into specific orthogonal space for regularizing so that the human action viewpoint can be normalized. Rogez et al. [18] estimated the 3D principal directions of man-made environments and the direction of motion, then transformed both 2D-Model and input images to a common frontal view before the fitting process. Though this kind of approaches can remove the viewpoint effect directly, the recognition results completely depend on the robustness of the body orientation estimation. Furthermore, the computational cost is significantly high.

The approach based on the state space models, e.g. Hidden Markov Model(HMM) and their extensions have been widely used for view-invariant action recognition. There are usually two kinds of state-space recognition methods to solve the viewpoint problem. The former approach directly recover the 3D human model by using tracking technology or pose estimation from the motion sequence. Then these 3D human model representations are utiliazed as the state information of the graph model for action recognizing [11, 16, 19]. Lv and Nevatia [11] exploit a large number of HMMs to model 3D human joints. And each HMM corresponds to the motion of a single joint or combination of related multiple joints. Peursum et al. [16] propose a variant of the hierarchical HMMs to achieve human action modeling and human body tracking simultaneously. These approaches exploit 3D information to represent the human model pose, so the viewpoint effect on human model is easily removed. However, inferring 3D poses from a single view usually is slow due to the large number of parameters that need to be estimated and it is still a non-trivial task to accurately detect and track the body ports in unrestricted scenarios. The latter approach based on state-space usually builds a multi-view posture image dataset of the human actions and exploits 2D contour information from different viewpoints to represent the same human static posture. Then the constraints on transition of synthetic poses and different viewpoints

are represented by a graphical model [12, 13, 15]. Lv et al. [12] introduced Action Net where each node stores a 2D representation of a 3D pose from a specific viewpoint. In the recognition phase, silhouette image from each frame is matched against all the nodes in the model using some distance measure. This method adopts contour information to represent silhouette image, so the accuracy of the recognition result strongly depends on the viewpoint similarity between the input images and training set. So a large number of training data from multiple cameras is needed to cover the possible camera viewpoint.

How to extract discriminative and robust features to describe actions and design new effective learning methods to fuse different types of features have become two important solutions for view-invariant action recognition. Because of the disadvantage of the above methods, a kind of method combining ideas from the state space method and template based threads was proposed to solve the viewpoint constrain issue. Ahmad and Lee [1] extracted the Cartesian component of optical flow velocity and human body silhouette feature vector information. Then they represented each action using a set of HMMs and classified a given sequence in any viewing direction by employing a likelihood measure. This method provides a novel solution to view-invariant human action recognition. However the method utilized the silhouette feature which is difficult to extract in the complex background. Furthermore the method didn't consider the probability fusion of the HMMs in different viewpoints. In our previous work, a view-insensitive feature combining the BoW of interest point in shot length-based video and the grid-based amplitude histogram of optical flow are used for representing the human motion information. The view space is partitioned into multiple sub-view space according to the camera rotation viewpoint. Human action models are trained by HMMs algorithm in each sub-view space. Finally the action with unknown viewpoint is recognized via the probability weighted combination [6]. This method obtained view-invariant human action reocgnition. However the method did not consider the viewpoint transition relationship, can not accurately recognize the actions captured from the top cameras. On the basis this work, a novel graphical representation i.e. multi-view transition HMMs by using combined feature [6] is proposed in this paper for view-invariant human action recognition. In the training phase, the view space is partitioned into finite sub-view spaces according to the rotation viewpoint. Then the sub-HMMs corresponding to human action models in a particular sub-view space are trained by using combined features. The combined features are view-insensitive, so the sub-HMMs can achieve better recognition performance to human actions with a certain range viewpoint changing. Finally the multi-view transition HMMs are built by connetcting the sub-view HMMs between adjacent viewpoint spaces. During the recognition process, the observation probability between the human action with unknown viewpoint and each trained action model by using improved forward algorithm. The action can be recognized corresponding to the maximum likelihood.The advantage of the proposed method does not require human body joint detection and strict temporal alignment.

2 The overview of the proposed method

The framework of the approach is sketched in Fig. 1, in which each human action is modeled by a multi-view transition HMM. The proposed multi-view transition HMM is built by connecting the sub models which are the action models in the particular viewpoint space. The framework includes the following modules:

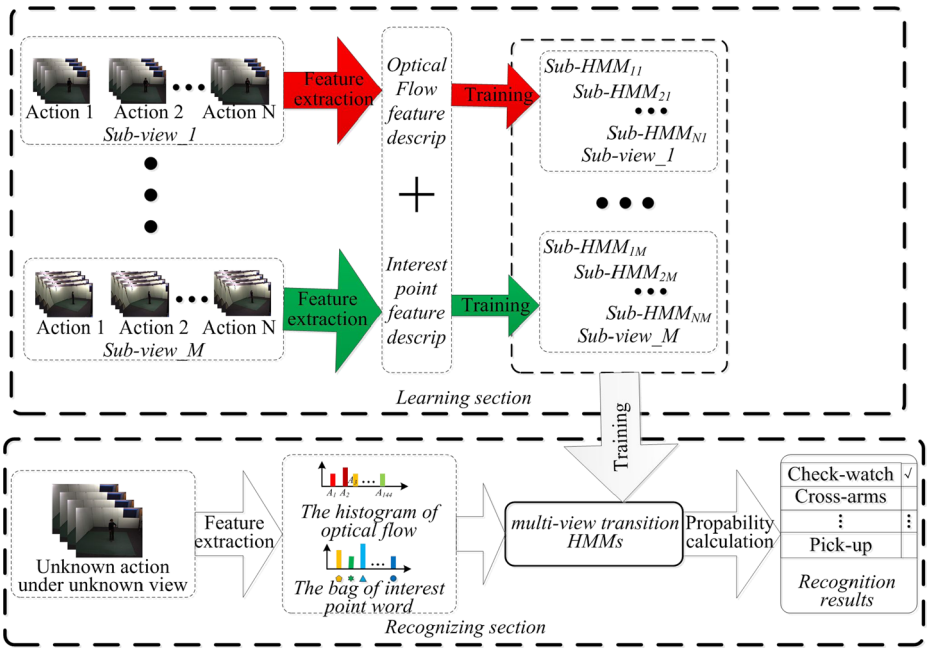

Fig. 1 The action recognition framework of the proposed approach

(1) **View Space Partitioning:** In order to improve the accuracy of view-invariant action recognition, the parameter space is decomposed into multiple sub spaces according to viewpoint. The actor's rotation viewpoint around the camera is equally partitioned into V pieces. Each piece will be regarded as a sub-view space. The training data in the same sub-view space can be collected by using viewpoint clustering. So it is not necessary to synchronously collect the training data by using multiple cameras in the proposed method.

(2) **Feature Extraction:** The robustness of the feature to viewpoint changes is helpful to improve the performance of recognition method. The combined features of the BoW of interest point in shot length-based video and the grid-based amplitude histogram of optical flow are selected to represent the human motion. The combined features are robust to the viewpoint change in the sub-view space.

(3) **Model Training:** The HMM is chosen as the sub-model to model the human action in the particular sub-view space. Firstly the sub-model of human actions are built by using the combined features which have been extracted from each sub-view space(each sub-view space contains H action models). Then the multi-view transition HMM is built by connecting the sub models with the viewpoint transition constraint, i.e. one multi-view transition HMM is according to one human action model.

(4) **Action Recognition:** When the test action sequence with unknown viewpoint is given, feature extraction is performed. Then the probabilities of the test sequence for the multi-view transition HMMs are computed by using the improved forward algorithm. The test action is recognized by using the maximum likelihood criterion.

3 Action feature

Single feature is weak to recognize the complex human actions. Different human action features have various discriminative abilities. In order to improve the robustness of the method to viewpoint changes, a view-insensitive motion feature combined the BoW description of interest point and the grid-based amplitude histogram of optical flow is utilized. The combined feature has been verified to be insensitive for viewpoint changing in our previous work [6]. The detail of the feature is shown as following.

3.1 Interest point feature extraction and description

Interest point feature do not require foreground segmentation or body tracking, so they are more robust to camera movement and low resolution. Approaches based on interest point have shown much success in action recognition [3, 14]. The conventional BoW description of interest point is usually extracted throughout the whole video. So it usually lacks the dynamic information in temporal domain. To address this issue, the BoW description in shot length-based video is proposed to improve the temporal characteristic and the contextual semantic. The BoW feature extraction and description process is shown as Fig. 2.

The detailed of extraction is outlined in the following:

(1) **Interest point detection:** The frames in the training dataset which contain rich motion information are found by interest point detection algorithm [3]. The method firstly detect the region of interest by the frame differencing algorithm, then filter the region by using 2D Gabor filters from different orientation (0°, 22°, 45°, 67° and 90° orientations selected). The combination of different orientation filtering responses is used for the final results of interest point detection.

(2) **3D SIFT description:** The 3D SIFT feature is verified it can capture rich local motion features and is robust to minor variations of viewpoint [6]. Then the interest points in these frames are represented by using the 3D Scale-invariant Feature Transform (3D SIFT) descriptor.

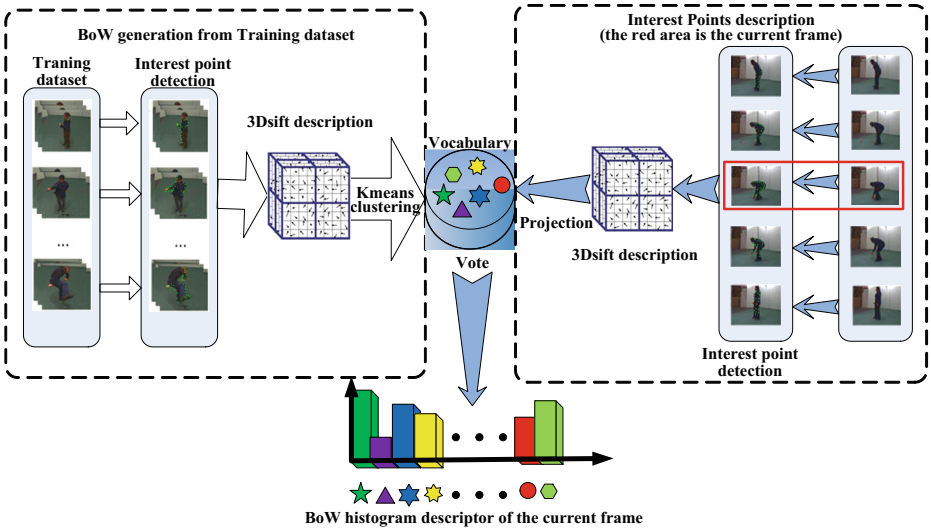

Fig. 2 The BoW feature extraction and description of interest point in shot length-based video

(3) **Vocabulary construction:** The 3D SIFT descriptors of the interest points captured from the training dataset are clustered by using K-means algorithm, and the K clustering centers are chosen as the words to construct the vocabulary.

(4) **BoW feature description:** When the new frame of test sequence is input, the interest points are extracted in the F neighbor frames to represent the current frame feature. The 3DSIFT descriptors of the extracted points are projected into the K-dimension vocabulary by minimizing the Euclidean distance between the descriptors of current neighbor frames and the words in vocabulary. Finally, the occurrence frequencies of the words are counted as the BoW descriptor of interest point for the current frame.

In order to investigate the effect of the different values of K and F to the system performance, the algorithm was tested with $K = 30, 45, 75, 100, 144$ and $F = 2, 3, 4, 5$. The results have shown that the combined feature reached the best performance with $K = 60$ and $F = 4$. So each frame can be represented by a 60-dimension BoW histogram descriptor.

3.2 Optical flow feature extraction and description

The BoW description of interest point in shot length-based video can effectively describe the local motion information, and the optical flow contains more motion information and is more robust to the complex environments. So the optical flow feature is chosen to combine with interest point feature. In order to reduce the computation complexity of optical flow feature, the optical flow features are extracted from the frames with rich motion information found by the interest point detection. The optical flow feature extraction and description process is shown as Fig. 3.

The detail is outlined in the following:

(1) **Optical flow calculation:** Optical flow images for horizontal and vertical channels are calculated in the regions of interest(ROI) of the adjacent gray frames by using Lucas-Kanade algorithm.

(2) **Normalization:** The sizes of ROI in every frame is different, so the normalization is utilized to scale the interest region into a fixed $p \times q$ ($p = q = 120$ in this paper) dimension by employing the bilinear interpolation.

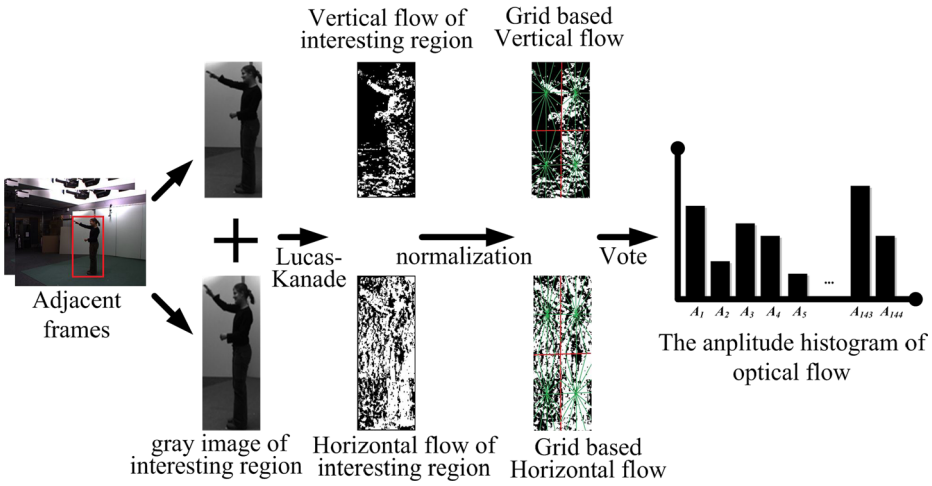

Fig. 3 Grid-based amplitude histogram of optical flow feature extraction

(3) **Optical flow feature description:** In order to increase the anti-disturbance ability, the grid-based method [20] is adopted to divide each normalized optical flow image into 2×2 sub-windows. Then a M-dimension radial histogram of the optical flow amplitude is computed by calculating optical flow amplitude in M pie slices ($M = 18$ in this paper). Finally, each frame can be represented by a 144-dimension ($2 \times 2 \times 2 \times 18$) optical flow feature.

To make the proposed approach more robust to complex environments and viewpoint changing, the BoW descriptor of interest point in shot length-based video are combined with the grid-based amplitude histogram of optical flow. So each frame will be represented by a 204-dimension ($144 + 60$) mixed feature. The combined features can not only capture the global motion information, but also achieve robustness to occlusions and viewpoint changes.

4 Multi-view transition hidden Markov model training and inference

In order to obtain the view-invariant human action recognition, the view space according to the rotation viewpoint is partitioned into V (V = 8, in this paper) sub-view spaces, as shown in Fig. 4. Only the actor's rotation viewpoint relative to the camera is considered. The variation range in a sub-view space is fixed(45° in this paper). For example, the viewpoint space 1 covers rotation viewpoint from −22.5° to 22.5°. A novel action model based on multi-view transition HMM is proposed to represent the human action. The model can not only recognize the human action with viewpoint changes, but also simplify the model training process by decomposing the parameter space into multiple sub-space.

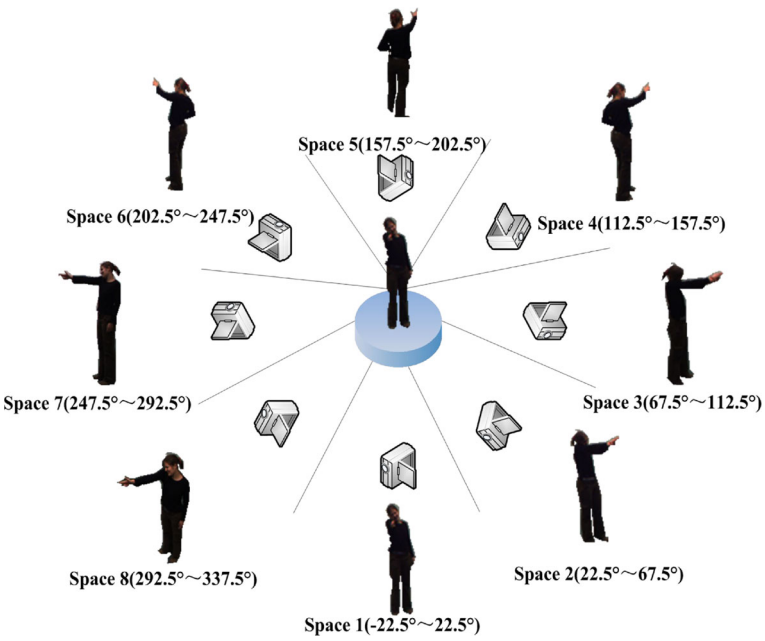

Fig. 4 Graphical representation of multiple viewpoint partition

4.1 Multi-view transition hidden Markov model

An example of the proposed action model is shown in Fig. 5, which corresponds to one human action model with different viewpoint transition. v and S respectively represent the current sub-view space and hidden state in the model. The sub model in the red box corresponds to a standard action HMM in a particular viewpoint space. There are three hidden states in the figure, however the number of the hidden states N maybe more than three in the experiment. Considering the change of actor's orientation should be smooth, i.e. the camera viewpoint should remain constant or change between adjacent viewpoints. So the viewpoint transitions in the graphical model only exist between the adjacent viewpoints.

4.2 Multi-view transition hidden Markov model training

The proposed action model is composed of 8 sub models, which is corresponds to a standard action HMM in a particular viewpoint space. So the proposed model decomposes the parameter space into 8 sub spaces which correspond to 8 sub-view spaces. So the parameter training in multi-view transition HMM can be divided into two parts: one is the parameter training of sub model in particular viewpoint, the other part is the state transition probability between adjacent viewpoints.

4.2.1 Sub-view space HMM training

Supposing that $\lambda^{(1,v)}, \lambda^{(2,v)}, \cdots, \lambda^{(H,v)}$are the trained HMMs of action1, action2,. . ., actionH in the current v sub-view space. The sub-view space HMM can be characterized by the following parameters:

1. N :The number of states in HMM (the number of hidden states). N states are denote as $S = \{s_1, s_2, \ldots, s_N\}$and the hidden state at time t as $q_t \in S = \{s_1, s_2, \ldots, s_T\}$.
2. T : The number of observation symbol in the sequence. The observation symbol sequence is denoted as $O = \{o_1, o_2, \ldots, o_N\}$.
3. A : State transition matrix $A = (a_{ij})_{N*N}$, where $a_{ij} = p(q_{t+1} = s_j | q_t = s_i)(1 \leq i, j \leq N)$?? a_{ij} is the probability of reaching state s_j at time $t+1$ from state s_i at time t.
4. B : Observation symbol probability distribution, $B = \{b_i(o_t)\}$, where $b_i(o_t) = p(o_t|s_i)(1 \leq i \leq N)$, $b_i(o_t)$ is the probability of generating observation symbol o_t from state s_i at time t.
5. π : The initial state distribution $\pi = \{\pi_1, \pi_2, \ldots, \pi_N\}$, where π_i is the probability of initial state s_i.

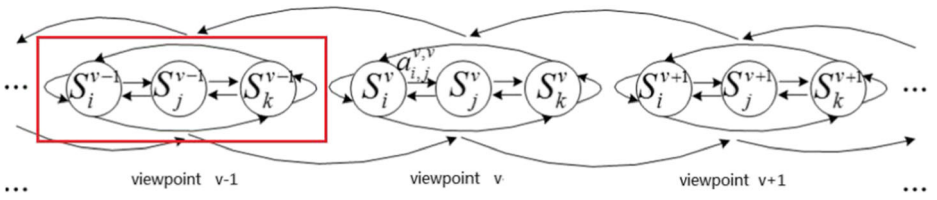

Fig. 5 Graphical structure example of the proposed HMM with multi-view transition

We denote a HMM as $\lambda = \{A, B, \pi\}$ using the above parameters. In doubly embedded stochastic process, parameter π, A describe the Markov chain and B describes the relation between state and observation symbol respectively.

The probability of generating observation symbol from each state can be computed by Gaussian probability-density function (1):

$$b_i(o_t) = b_{(u_i,\Sigma_i)}(o_t) = \frac{1}{\sqrt{2\pi}^d \sqrt{|\sum_i|}} e^{-\frac{1}{2}(o_t-u_i)^T \sum_i^{-1}(o_t-u_i)} \tag{1}$$

Where $u_i, \sum_i$ is respectively the mean and covariance matrix of observations in cluster i; d is the dimension of observation symbol o_t; $(o_t - u_i)^T$ is the transpose of matrix $(o_t - u_i)$; $\sum_i^{-1}$ is the inverse of matrix $\sum_i$.

The essence of HMMs training problem with the given structure (5 states and full-connected HMMs in this paper) is to maximize the observation probability by adjusting the model parameter for the observation sequence. Baum-Welch algorithm is commonly used to obtain the optimal model parameters. However, it's performance is depended on the choice of initial parameter. If the improper initial parameters are chosen, it can lead procedure to the local minimum. So the action model can't be optimal. In this paper the result of kmeans algorithm is taken as initial input of the Baum-Welch algorithm. Then the Gaussian probability density function is used for computing the probability of generating observation symbol from each state.

4.2.2 State transition probability calculation between adjacent viewpoints

The viewpoint constrains should be considered before calculating the transition probability. The state transition probability only exists between adjacent viewpoints. We assume a uniform transition probability between adjacent viewpoints, i.e., the state in viewpoint space v has the uniform transition probability to viewpoint space $v+1$, viewpoint space v and viewpoint space $v-1$. Considering there are N states in each HMM, the current state can be transformed to $3*N$ state nodes. So the transition probability $a_{i,j}^{m,n} = p(q_{t+1} = S_j^n | q_t = S_i^m)$, $1 \le i, j \le N$ can be calculated by (2).

$$a_{i,j}^{m,n} = \frac{1}{3N}, m \ne n, 1 \le i, j \le N \tag{2}$$

There m, n is the number of the sub-view space.

Then the multi-view transition HMM can be built by combining the parameters of sub-view space action HMM and the state transition probability between sub-view space. The state transition matrix can be described by (3).

$$A = (a_{i,j}^{m,n})_{VN \times VN}, a_{i,j}^{m,n} = \begin{cases} \frac{1}{3} a_{i,j}^m, m = n \\ \frac{1}{3N}, m \ne n \\ 0, others \end{cases} \tag{3}$$

In the current viewpoint space, the probability of generating observation symbol o_t from state s_i^v at time t, $b_i^v(o_t) = p(o_t | s_i^v)$ can be calculated by using (1).

The initial state distribution, $\pi = \{\pi_1^1, \pi_2^1, \ldots, \pi_N^1 \ldots \pi_1^v, \pi_2^v, \ldots, \pi_N^v\}$, where $\pi_i^v = \frac{1}{V} \times \pi^{(v)}$ is the probability of initial state s_i^v.

4.3 Multi-view transition hidden Markov model inference

Supposing that $\lambda^{(1)}, \lambda^{(2)}, \cdots, \lambda^{(H)}$ are the trained multi-view transition HMMs of action1, action2,. . ., actionH. When a test observation sequence $O = \left\{o_1^{test}, o_2^{test}, \cdots o_T^{test}\right\}$ is given, the probability of the test sequence (i.e. observation probability) for the given multi-view transition HMMs can not be computed by using traditional Forward-Backward algorithm. So the forward algorithm is improved to calculate the observation probability, as shown in Fig. 6

The forward probability is redefined as: $\alpha_t(v, i) = p(o_1, o_2, \ldots, o_t, q_t = S_i^v|\lambda)$, it is the probability of being in the state S_i^v at time t and having observed the sequence $O = \left\{o_1^{test}, o_2^{test}, \cdots o_T^{test}\right\}$ under the given action model.

Then the observation probability of the test sequence for the given multi-view transition HMMs can be calculated by following iterative formula.

1. **Initialization:** $\alpha_t(v, i) = \pi_i^v b_i^v(o_t)$
2. **Loop:** $\alpha_{t+1}(v, j) = \sum_{m=v-1}^{v+1} \sum_{i=1}^{|S^m|} \alpha_{i,j}^{m,v} b_j^v(o_t)$
3. **End:** $p(O|\lambda) = \sum_{m=1}^{V} \sum_{i=1}^{|S^m|} \alpha_T(m, i)$

The observation probability $p(O|\lambda^{(1)}), p(O|\lambda^{(2)}), \ldots, p(O|\lambda^{(H)})$ between the test sequence and each trained action model are computed by using above improved forward

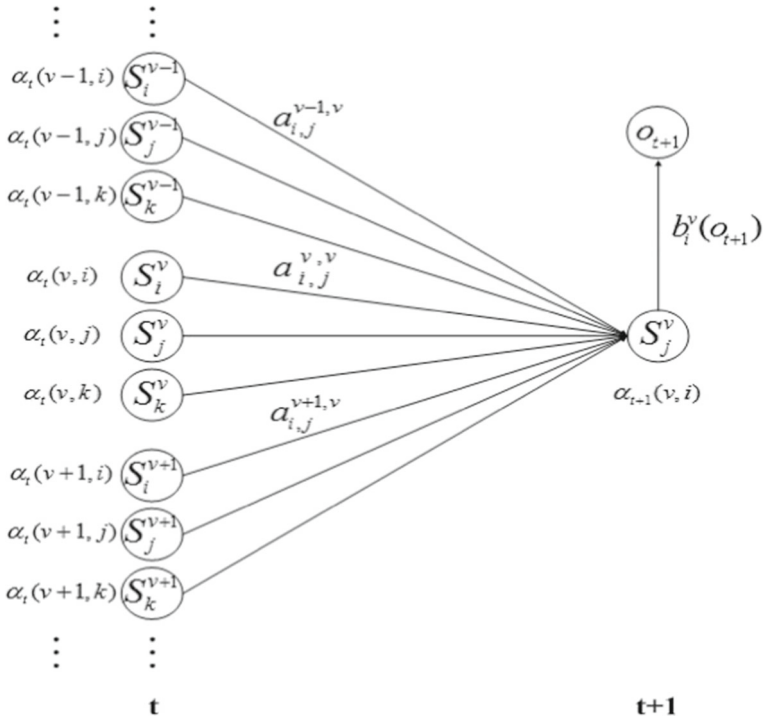

Fig. 6 The iterative process of Forward algorithm with multi-view transition

algorithm. Then the action corresponding to the maximum likelihood is chosen as the best recognition action:

$$test\_number = \arg\max \sum_{1 \le h \le H} (p(O|\lambda^{(h)})) \tag{4}$$

5 Experiment

The proposed method is tested on the Inria Xmas Motion Acquisition Sequences (IXMAS) multi-view action dataset, which contains twelve daily-live actions. Each action is performed three times by twelve actors and recorded simultaneously from five different cameras: four side cameras and one top camera. During the action performing, the viewpoint of human body relative to each camera is not restricted. The examplar frames are shown as Fig. 7. Whether under the same camera or different cameras there are large viewpoint variation of the human body relative to the camera. Therefore, it is widely used to test the performance of view-invariant action recognition algorithms. Eleven actions are chosen for verifying our approach during the experiment, namely, check-watch, cross-arms, scratch-head, sit-down, get-up, turn-around, walk, wave, punch, kick and pick-up. So there are $5 * 11 * 3 * 12$ action samples for the experiments. The test method of leave one actor out (LOAO) is used in every experiment.

5.1 View clustering

The multi-view IXMAS dataset was captured by using five cameras placed different positions around the actors. In this dataset, actor orientations are arbitrary since no specific instruction is given during the acquisition. That is to say, although the actions are captured by the same camera, the camera viewpoints are different when the actions are performed by different actors.

In order to train the sub action models under a particular camera viewpoint, the action sequences were clustered into 8 different viewpoint spaces depending on the rotation viewpoint, as shown in Fig. 4. Since there are not enough samples captured from viewpoint 5

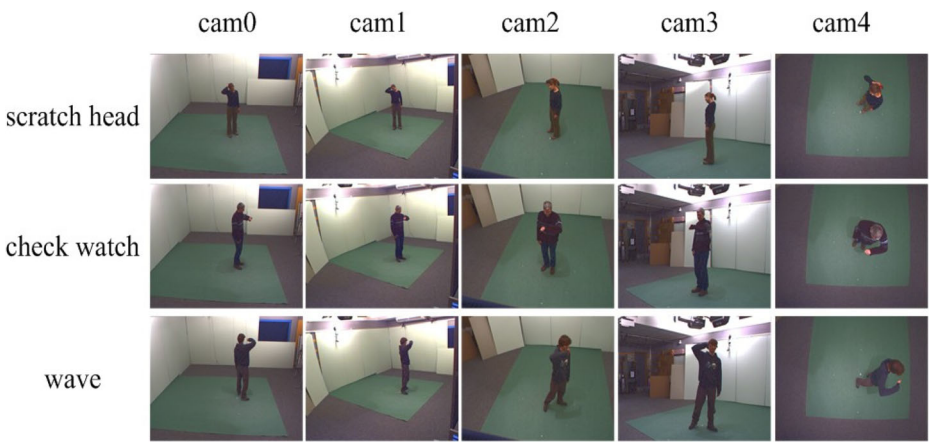

Fig. 7 Exemplar frames from IXMAS multi-view dataset

and 6, the action samples from those two viewpoint spaces are not included to train the sub action models. That is to say, only 6 viewpoint spaces is considered in this paper.

5.2 The view-invariant action recognition test under particular camera

The viewpoint variation of the human body relative to the camera can be divided into two types: rotation variation and pitch variation. In the dataset, the pitch viewpoint variation of the human body is relatively small. In this experiment, the performance of the multi-view transition HMMs for view-invariant action recognition under particular camera is tested. The experimental data respectively used for model training and testing are taken from the same camera data(the testing actor sequences are not included in the training data). Firstly 6×11 sub-action models are trained according to 6 viewpoint spaces and 11 class of human action under the same camera. Then the multi-view transition HMMs are built by connecting the sub-action models between the adjacent viewpoint spaces. The performance of the proposed method to the rotation viewpoint changing can be verified in this experiment. The recognition rates of every action are listed in Table 1.

The results show that the proposed method is extraordinary effective to discriminating the human actions with viewpoint changing and achieve a satisfying recognition rate. Especially the algorithm for the top camera can obtain 86.6 % recognition rate. The results is very inspiring, since it is very difficult to recognize the human actions from the top camera.

5.3 The view-invariant action recognition test under different cameras

On the basis of the above experiment, the performance of the proposed method for the real view-invariant recognition (rotation-invariant and pitch-invariant) is considered in this experiment. The data from different cameras (four side cameras, one top camera) are used to train the multi-view transition HMMs. There exists large patch-variation between different cameras. The view space is partitioned as usual according to the rotation viewpoint of human body to the camera. That is to say the data with the same rotation viewpoint will be clustered into the same sub-view space, no matter which camera the data belong to. At each test iteration, the action samples performed by one actors from one camera are chosen as

Table 1 The recognition results (%) of multi-view transition HMM under particular camera

parameter	cam0	cam1	cam2	cam3	cam4
check-watch	80	96	88	92	88
cross-arms	88	96	84	84	80
scratch-head	72	76	88	68	92
sit-down	96	100	100	96	92
get up	96	100	100	96	80
turn-around	88	80	80	88	92
walk	100	100	100	100	100
wave	80	68	88	92	84
punch	100	92	92	88	76
kick	100	96	88	80	80
pick-up	100	100	100	96	88
average	91.3	91.3	91.6	89.1	86.6

Table 2 The recognition results (%) of multi-view transition HMM under particular camera

parameter	cam0	cam1	cam2	cam3	cam4	average
Recognition Rate	91.6	92.7	86.9	84.8	76.0	86.4

testing sample, then multi-view transition HMMs are learned by using the remaining training samples performed by other actors from the same camera and all other samples from other cameras. Totally only 11 multi-view transition HMMs are trained to achieve the view invariant human action recognition under different cameras. The iteration process will be terminated until all the samples are tested. The recognition rates of every camera are calculated and listed in Table 2. The confusion matrix of whole testing is provided in Fig. 8 to show the effectiveness of the proposed method.

Observed recognition results from Table 2, the proposed multi-view transition HMMs based action recognition method can achieve better recognition performance for the action data captured from different cameras. It is verified that the proposed method has better robustness to viewpoint changing (no matter rotation viewpoint and pitch viewpoint) with less training model. Only one multi-view transition HMMs is trained to model one class of human action in the whole viewpoint spaces.It is worth noting that the proposed method achieve better recognition results by using action data from camera 4, although there exists large pitch viewpoint variation according to other cameras.

Furthermore, we can find the proposed method can successfully recognize most of the test data. The wrong recognition results are mainly focus on the similar actions of "scratch head" and "wave". The main reason is the recognition of those two class of actions depending on the information captured from the arm of the actors. However there occur overlapping phenomenon in the most of the duration of action performing, as shown in Fig. 9. So there are not enough discriminative information obtained from the image sequence, which leads to false recognition between "scratch head" and "wave".

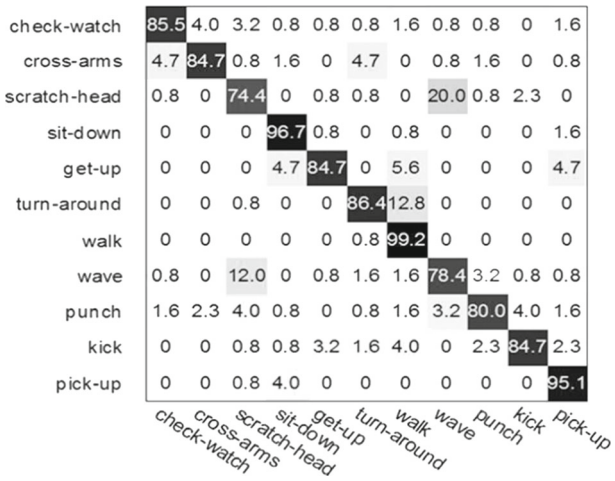

Fig. 8 Confusion matrixes (in %) for the IXMAS dataset

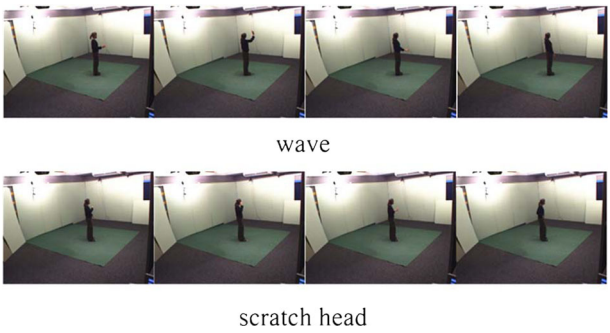

Fig. 9 Confusion matrixes (in %) for the IXMAS dataset

5.4 Performance comparison

The proposed approach is compared with the state-of-the-art view-invariant methods. To make the comparison, our experiment is performed under the same test conditions with them. All of the videos in the dataset are used as experimental data. The leave one actor out (LOAO) is used as standard testing method. And the performance for different recognition methods is illustrated as Table 3.

As can be seen from Table 3, the work by Liu et al. [10] greatly outperforms than our method. However this result is obtained by fusing five cameras inputs. The average recognition rate achieved by using one camera data is 82.38 %. It is lower than our method. Furthermore, the recognition rate of our previous work [6] is slightly better than our current work. However this result is the average recognition rate of four cameras, and top cameras is not included. So under the same testing condition, it is worth noting that our method outperforms all of other methods. During feature extraction process, neither the key poses [12] are extracted from motion capture sequences nor the camera information [10] are required in advance. By effectively introducing the viewpoint transition, the recognition accuracy and the robustness for viewpoint changing are improved. The experimental results show that the proposed approach is easy to implement and the performance for view-invariant action recognition is satisfactory.

Table 3 Comparison with related work in recent years

Literature	Method	Accuracy
Wu et al. [23]	ST context + appearance + AFMKL	78.02 %
Lv et al. [12]	Shape context of silhouettes +PMK-NUP	80.6 %
Junejo et al. [7]	SSM desciptor +SVM	72.7 %
Weiland et al. [22]	3D HOG descriptor +Local SVM	83.4 %
Liu et al. [10]	Silhouette -optical flow -interest point + LWE	82.8 %
Wu et al. [24]	Correlogram of body poses + Multi-Max-Margin SVM	95.54 %
Our previous work [6]	Optical flow+interest point SIFT + multi-HMMs probability fusion	88.4 %
Our approach	Optical flow+interest point SIFT + multi-view transition HMM	86.4 %

6 Conclusion

A novel view-invariant human action method based on multi-view transition HMMs is proposed in this paper. The interest point feature based on local information and the optical flow feature based on global motion information are effectively combined. The combined feature is view-insensitive in the sub-view space. The multi-view transition HMMs are built in the training process by introducing the transition probability matrix to achieve the transition of sub-action models in multi-view space. The sub-action models can be independently trained, which can greatly reduce the computational complexity. Finally the action with unknown viewpoint is recognized by using improved forward algorithm. The experiments validated the performance of recognition algorithms is greatly improved and it is superior to the existing view-invariant action recognition algorithms. With the use of the RGBD cameras, it is our future work to solve the viewpoint constrain issue by fusing depth information [2, 8].

Acknowledgments The Project supported by the National Natural Science Foundation of China No. 61103123 and the Program for Liaoning Excellent Talents in University (No. LJQ2014018).

References

1. Ahmad M., Lee S. (2006) Hmm-based human action recognition using multiview image sequences. In: the 18th IEEE int. conf. pattern recognition (ICPR), pp 263–266
2. Ashraf N, Sun C, Foroosh H (2014) View invariant action recognition using projective depth. Int J Comput Vis Image Underst 123:41–52
3. Bregonzio M, Gong S, Xiang T (2009) Recognising action as clouds of space-time interest points. In: the 27th IEEE conf. computer vision and pattern recognition (CVPR). USA, pp 1948–1955
4. Holte M, Moeslund T (2011) Human action recognition using multiple views: a comparative perspective on recent developments. In: the 2011 joint ACM workshop on human gesture and behavior understanding, pp 47–52
5. Ji X, Liu H (2010) Advances in view-invariant human motion analysis: a review. IEEE Trans Syst Man Cybern Part C 1:13–24
6. Ji X, Wang C, Li Y (2014) A view-invariant action recognition based on multi-view space hidden markov models. Int J Humanoid Rob 11(1):1–17
7. Junejo I, Dexter E, Laptev I, Perez P (2008) Cross-view action recognition from temporal self-similarities. In: the 10th European conf. computer vision. France, pp 293–306
8. Lee AR, Suk HI, Lee SW (2014) View-invariant 3D action recognition using spatiotemporal self-similarities from depth camera. In: the 22nd int. conf. pattern recognition (ICPR). Sweden, pp 501–505
9. Lim C, Vats E, Chan C (2015) Fuzzy human motion analysis: a review. Pattern Recogn 48(5):1773–1796
10. Liu J, Shah M, Kuipers B, Savarese S (2011) Cross-view action recognition via view knowledge transfer. In: the 29th IEEE conf. computer vision and pattern recognition (CVPR). USA, pp 3209–3216
11. Lv F, Nevatia R (2006) Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In: the 9th European conf. computer vision (ECCV). Austria, pp 359–372
12. Lv F, Nevatia R (2007) Single view human action recognition using key pose matching and Viterbi path searching. In: the 25th IEEE int. conf. computer vision and pattern recognition (CVPR). USA, pp 1–8
13. Natarajan P, Nevatia R (2008) View and scale invariant action recognition using multiview shape-flow models. In: the 26th IEEE int. conf. computer vision and pattern recognition (CVPR), pp 1–8
14. Niebles JC, Wang H, Feifei L (2008) Unsupervised learnling of human action categories using spatial-temporal words. Int J Comput Vis 79:299–318
15. Peng B, Qian G, Rajko S (2008) View-invariant full-body gesture recognition from video. In: the 19th int. conf. pattern recognition (ICPR). USA, pp 1–5
16. Peursum P, Venkatesh S, West G (2007) Tracking as recognition for articulated full body human motion analysis. In: the 25th IEEE int. conf. computer vision and pattern recognition (CVPR). USA, pp 1–8
17. Poppe R (2010) A survey on vision-based human action recognition. Image Vis Comput 28(6):976–990

18. Rogez G, Guerrero JJ, Martinez J, Orrite C (2006) Viewpoint independent human motion analysis in man-made environments. In: the 17th British machine vision conference (BWVC). UK, pp 659–668
19. Singh VK, Nevatia R (2011) Simultaneous tracking and action recognition for single actor human actions. IEEE Trans Vis Comput Graph 27(12):1115–1123
20. Tran D, Sorokin A (2008) Human activity recognition with metric learning. In: the 10th European conf. computer vision (ECCV). France, pp 548–561
21. Weinland D, Ronfard R, Boyer E (2006) Free viewpoint action recognition using motion history volumes. Comput Vis Image Underst 104(2):249–257
22. Weinland D, Ozuysal M, Fua P (2010) Making action recognition robust to occlusions and viewpoint changes. In: the 10th European conf. computer vision. Greece, pp 635–648
23. Wu X, Xu D, Duan L, Luo J (2011) Action recognition using context and appearance distribution features. In: IEEE int. conf. on computer vision and pattern recognition (CVPR). USA, pp 489–496
24. Wu D, Shao L (2014) Multi-max-margin support vector machine for multi-source human action recognition. Neurocomputing 127:98–103
25. Yan Y, Ricci E, Subramanian R, Liu GW, Sebe N (2014) Multitask linear discriminant analysis for view invariant action recognition. IEEE Trans Image Process 23(12):5599–5611
26. Yilmaz A, Shah M (2005) Actions sketch: a novel action representation. In: the 23th IEEE int. conf. computer vision and pattern recognition (CVPR). USA, pp 984–989

Xiaofei Ji received her M.S. and Ph.D. degrees from the Liaoning Shihua University and University of Portsmouth, in 2003 and 2010, respectively. From 2003 to 2012, she was the Lecturer at School of Automation of Shenyang Aerospace University. From 2013, she holds the position of Associate Professor at Shenyang Aerospace University. She is the IEEE member, has published over 40 technical research papers and 1 book. More than 20 research papers have been indexed by SCI/EI. Her research interests include vision analysis and pattern recognition. She is the leader of National Natural Science Fund Project(Number: 61103123) and main group member of 6 National and Local Government Projects.

Zhaojie Ju received the B.S. in automatic control and the M.S. in intelligent robotics both from Huazhong University of Science and Technology, China, in 2005 and 2007 respectively, and the Ph.D. degree in intelligent robotics at the University of Portsmouth, UK, in 2010. Dr Ju is currently a Lecturer in the School of Computing, University of Portsmouth, UK. He previously held research appointments in the Department of Computer Science, University College London and Intelligent Systems and Biomedical Robotics group, University of Portsmouth, UK. His research interests are in machine intelligence, robot learning, pattern recognition and their applications in robotic/prosthetic hand control and human-robot interaction.

Ce Wang received his B.Eng. degree in Automation Engineering from Harbin Institute of Technology, Weihai, China, in 2011. He is currently a graduate student studying for Master degree in the School of Automation, Shenyang Aerospace University. His research is focus on the human action modeling and recognition. He has published 4 research papers in this research direction.

Changhui Wang received his B.Eng. degree in Measurement and control technology and instrument from Yanshan University, Linren college, China, in 2013. He is currently a graduate student studying for Master degree in the school of Automation, Shenyang Aerospace University. His research is focus on the human interactive behavior recognition.