다중 시점 영상 시퀀스를 이용한 강인한 행동 인식

아마드^0, 이성환
고려대학교 정보통신대학 컴퓨터·통신공학부
{mohi, swlee}@image.korea.ac.kr

Robust Action Recognition Using Multiple View Image Sequences

Mohiuddin Ahmad^0, Seong-Whan Lee
Division of Computer and Communication Engineering, Korea University

요 약

Human action recognition is an active research area in computer vision. In this paper, we present a robust method for human action recognition by using combined information of human body shape and motion information with multiple views image sequence. The principal component analysis is used to extract the shape feature of human body and multiple block motion of the human body is used to extract the motion features of human. This combined information with multiple view sequences enhances the recognition of human action. We represent each action using a set of hidden Markov model and we model each action by multiple views. This characterizes the human action recognition from arbitrary view information. Several daily actions of elderly persons are modeled and tested by using this approach and they are correctly classified, which indicate the robustness of our method.

1. Introduction

This contribution addresses the human action recognition of elderly people from arbitrary direction using their combined features, such as motion features and shape feature during performing the action. Recognition of action from image sequences is very popular in computer vision community, since it has applications in video surveillance and monitoring, human-robot interactions, etc. Actually, there is no rigid syntax and well defined structure is available for human action recognition. This makes human action recognition a more challenging and sophisticated task.

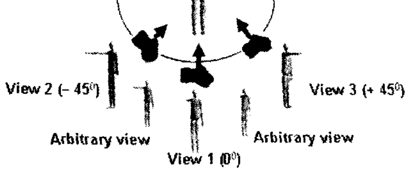


Figure 1. Human action in multiple view image sequences.

Several human action recognition methods have been proposed in the past decades. A recent detailed excellent survey can be found in [1]. Researchers either use the human body shape or silhouette information [10,12] or human motion information [8,11,13] for action recognition. Most of the above action recognition techniques depend on the viewing direction. The work of testing an action using multi-view motion learning is still unsolved. In [7,14], authors used the view-invariant approaches. There are some issues that affect the development of models of actions and classifications, which are as follows: (i) Action can be viewed by the motion of the human body parts, (ii) simple action and complex action involve the motion of small number and large number of body parts, respectively and the motion is non-rigid in nature, (iii) an action can be viewed as a series of silhouette images of the human body and silhouette information that involves no translation, rotation and scaling, (iv) same action from different viewing direction appears different and some part of the body may in occlusion, shown in Fig. 1.

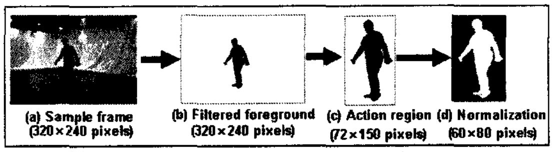
Based on the issues, motion of the body parts and human body silhouette play important role for recognition. Motion based feature can reveal the approximation of moving direction of human body and human action can be effectively characterized by motion rather than other cues, such as color, depth, and spatial features. On the other hand, human body silhouette represents the pose of the human body at any instant of time, and a series of silhouette images can be used to recognize human action successfully. Therefore, the combined feature can enhance the recognition accuracy of human actions. Moreover, the same action from different viewing angle looks different. Therefore, recognizing human action from multiple view sequence is a difficult task. We propose to model and recognize several actions of human using the combination of (i) optical flow vectors, (ii) shape feature vectors with multiple view image sequences. These characterize robust action recognition. The actions modeling and classification in this work are: walking at a place, raising a right hand, bowing, running at a place, and sitting on the floor, respectively. We use hidden Markov model for training each actions in any viewing direction. Classification is finally achieved by feeding a given (test) sequence in any viewing direction to all the trained HMMs and employing a likelihood measure to declare the action performed in the image sequence. For training and testing actions, we use the Korea University gesture database [5].

This paper is organized as follows: Section 2 briefly summarizes the foreground extraction algorithm. Section 3 describes the feature extraction procedures. Section 4 describes briefly the HMM for modeling and classifying action. Experimental results and discussions of the selected approaches are presented in Section 5. Finally, conclusions are drawn in Section 6.

2. Preprocessing

We use background subtraction to extract the foreground, since the background is relatively static for all over the image sequence of an action. Therefore, we adopt simple background modeling technique such as multiple Gaussian background modeling for foreground extraction. After that, there still exits some noises in the foreground, such as motion shadow will inescapably existed in the foreground. Therefore, shadow elimination method should be adopted. After the shadow elimination step, there may exist some small regions and noises, so several filters such as erosion, dilation, and connected component analysis should be adopted for further preprocessing. There still exist some holes or discrete pixels outside the original human body silhouette; these can be removed by using median filtering.

After preprocessing, we define an action region as the rectangular area or a bounding box where an action is occurred. The action region depends on the distance between sensor and persons, person's anthropometry, varieties of actions. Usually, action region is smaller than the image area; therefore, we select the action region inside the image frame which includes approximately the average human body shape of specific image sequence of the specified action. The action region of an image is extracted automatically from the filtered foreground image by using row-column scanning. The bounding box are for a typical action is 90×160 pixels. This varies according to action types.



(a) Sample frame (320×240 pixels) (b) Filtered foreground (320×240 pixels) (c) Action region (72×150 pixels) (d) Normalization (60×80 pixels)

Figure 2. Preprocessing steps.

Figure 2 shows the preprocessing steps, where the filtered foreground image (Fig.2(b)) results from background subtraction, shadow elimination, morphological operation and filtering. The action region (Fig.2(c)) is extracted from foreground. For a unique representation, we normalize the action region to a fixed size. The action region is normalized by using nearest neighbor (NN) interpolation method, shown in Fig. 2(d).

3. Feature Extraction

In the feature extraction procedure, we extract the motion features and shape features from the specified action region.

3.1 Shape features

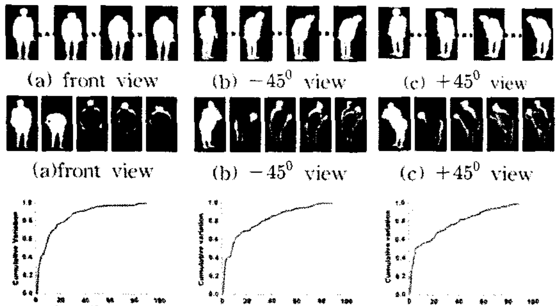
The silhouette images of the action region are normalized to p pixels by q pixels by using nearest neighbor interpolation method, shown in Fig. 3(d). Now, the human body silhouette may be considered as a vector of dimension pq so that a typical action region 60×80 becomes a vector of 4800, or a point of 4800-dimensional space. Thus, all the human body images in a sequence can form a sparse matrix which represents a higher dimensional feature space. For efficient use of the normalized action region silhouette (NARS) image, we use principal component analysis (PCA) to reduce the high dimensional feature space into a lower dimensional new feature space. PCA has been extensively used in the field of face recognition. The use of PCA in action recognition has been limited. In the paper, we use PCA to extract the silhouette feature vectors for pre-recognition of human action. Let the training set of the NARS images are$\Gamma_1, \Gamma_2, \ldots, \Gamma_M$. The average NARS of the set is defined by $\Psi = \frac{1}{M}\sum_{n=1}^{M}\Gamma_n$. Each NARS image differs from the average Ψ by the vector $\Phi_i = \Gamma_i - \Psi$. For any direction, this very large set of vectors is then subject to PCA, which seeks a set of M ortho-normal vectors, u_n, which best describe the distribution of the data. The$k$-th vector, u_k, is chosen [3], such that,

$$\lambda_k = \frac{1}{M}\sum_{n=1}^{M}(u_k^T\Phi_n)^2 \qquad (1)$$

The vectors, u_k and scalars λ_k are the eigenvectors and eigenvalues, respectively, of the covariance matrix of equation (2):

$$C = \frac{1}{M}\sum_{n=1}^{M}\Phi_n\Phi_n^T \qquad (2)$$

The eigenvectors, u_k, for $i = 1, 2, \ldots, M$ are ranked according to their associated eigenvalues, λ_k. Now, A new human body silhouette image (Γ) is transformed into its eigen-body silhouette components (projected into the human body silhouette space) by the equation, $s_k = u_k^T(\Gamma - \Psi)$ for $k = 1, 2, \ldots, R$.


(a) front view (b) -45^0 view (c) $+45^0$ view

(a)front view (b) -45^0 view (c) $+45^0$ view

(a) front view (b) -45^0 view (c) $+45^0$ view

Figure 3. Silhouette images (upper-row), corresponding eigen bowing images (middle row), and cumulative variation of feature vectors (lower-row).

We select the number of R new feature vectors so that maximum silhouette energy can preserve during learning the features. Therefore, we use new shape feature vectors, $F = [s_1, s_2, \ldots, s_R]^T$ for every frame of the sequence of any action in any viewing direction, Figure 3 shows the sample silhouette images (upper-row) and corresponding eigen-images (middle row), where the eigen-images for highest eigenvalues represent the maximum energy of silhouette, and cumulative variation of feature vectors are shown in Fig. 3 (lower-row).

3.2 Motion features

In this paper, we use optical flow to estimate motion, because it can precisely determine the motion at any pixel. We calculate the optical flow velocity at any pixel in the action region by using gradient based optical flow technique. The well known optical flow constraint equation [2] or gradient constraint equation is given by

$$p_x v_x(x,y,t) + p_y v_y(x,y,t) + p_t = 0 \qquad (3)$$

Here, $v_x(x,y,t)$ and $v_y(x,y,t)$ represent the horizontal and vertical component of optical flow motion, and p_x, p_y and p_t represent the horizontal gradient, vertical gradient, and temporal gradient, respectively. Equation (3) consists of two unknown components, constrained by only one linear equation. So, more constraints are required to determine the velocity components. We can consider the global motion of human, when he/she performs an action. We use the Horn's optical flow method [2] for calculating the motion components from consequent image frames.

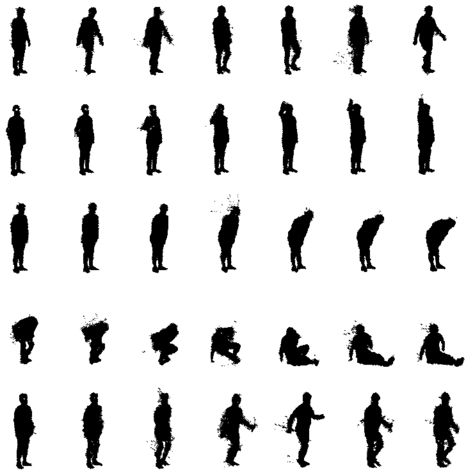


Figure 4. Global motion overlapping the human in action. Row-1: Walking, row-2: raise hand, row 3: bowing, row 4: sitting on floor, and row 5: running.

Figure 4 shows the optical flow velocity overlapping on the action region of several actions. It is found that the related body parts involve optical flow velocity. For example, when the person acts "raising a right hand" action, then motion involves only the right hand. Similarly, when the person acts "sitting on the floor" action, then motion involves the whole body.

For consistency of additional analysis, we normalize the optical flow values, $v_n(x,y,t)$ in the action region. In order to extract the features, we partition the action region into multiple blocks, $B(k)$ of equal sizes. Therefore, the optical flow feature vectors are extracted at each block with n number of pixels by using the equations (4a) & 4(b):

$$v_{kx,t} = \frac{1}{n}\sum_{(x,y)\epsilon B(k)} v_{nx}(x,y,t) \qquad (4a)$$

$$v_{ky,t} = \frac{1}{n}\sum_{(x,y)\epsilon B(k)} v_{ny}(x,y,t) \qquad (4b)$$

where, $k=1,2,\dots,B$ and B=no. of blocks. Figure 5 shows sample optical flow in different views with optical flow feature image.

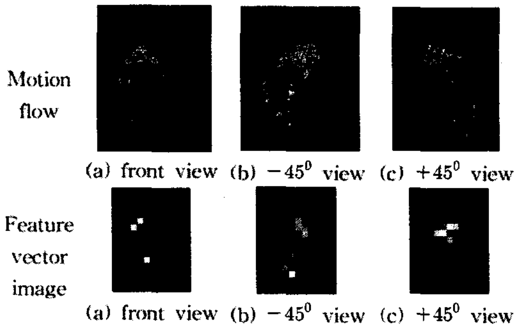

Motion flow
(a) front view (b) -45^0 view (c) $+45^0$ view
Feature vector image
(a) front view (b) -45^0 view (c) $+45^0$ view

Figure 5. Motion image with feature vectors image.

3.3 Combined features

For each image frame in any action, human body silhouette features and optical flow velocity features are combined as one feature vector. The features used at any instant of time, t of image sequence (per image frame) are as follows:

$$I_t = [v_{1x,t},\dots,v_{Lx,t},v_{1y,t},\dots,v_{Ly,t},s_{1,t},\dots,s_{M,t}]^T \qquad (5)$$

where, L represents the motions features and M represents the shape features. Each action video in any view direction d can be represented as an image sequence by the following equation (6):

$$H_d = [I_{1,d}, I_{2,d},\dots,I_{N,d}] \qquad (6)$$

where, N represents the number of frames used in an action video. The value of N depends on the action type and the performers; typically $N=60\sim200$. To use the body silhouette features and optical flow velocity features as the input of HMM, we convert the feature vector into discrete symbols. In our experiment, each feature vector is converted into one-dimensional sequence of 32 discrete symbols.

4. Action Modeling & Classification by HMMs

Hidden Markov Models has been successfully used for speech recognition. We employ HMM for action recognition since it can be applied to analyze the time series with spatio-temporal variations.

4.1 Action modeling

In this paper, we use multiple observations HMMs for modeling human actions. For multi-view recognition of human actions, we build HMM model for each action and for each viewing direction. For a given action, we use HMM model $\lambda_{ad}=\{A,B,\pi\}$ for any model action a in any viewing direction d, where $A,B,$ and π are defined in [9]. We can consider a set of hidden Markov models for the multi-view directions which is expressed as:

$$\lambda_a = \{\lambda_{a1},\lambda_{a2},\dots,\lambda_{ad}\} \qquad (7)$$

Each model represents the action from a specified viewing angle. We use same topology for all HMMs, i.e. we use six-state ergodic models. The number of states is heuristically selected. Since, we use the motion feature vector and the shape feature vector for action recognition, then we consider multiple observable symbols, O at each time t. We model multiple observation using discrete HMM models. We use Baum-Welch algorithm [4][9] for iteratively re-estimate model parameters to achieve the local maximum.

4.2 Action classification

We classified the image sequences manually into different classes and different views. The trained model is used to classify the actions. The forward-backward algorithm or the Viterbi algorithm can be used to classify the actions from any specified view. The model parameters are adjusted such a way that they can maximize the likelihood function for classifying actions by using the given set of training data.

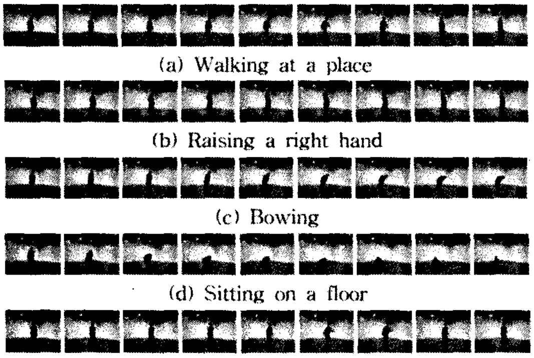
$$\lambda = \arg\max_{\lambda_a \epsilon\, actions} P(O|\lambda_a) \qquad (8)$$

where, $P(O|\lambda_a)$ is the conditional probability for any arbitrary action a and it is computed by $P(O|\lambda_a) = \max P(O|\lambda_{ad})$, where, O is the observations feature vector sequence of an unknown action. For the observation sequence, $O = [O^{(1)}, O^{(2)}, ..., O^{(T)}]$ and the HMM λ_{ad}, according to Bayes rule, the problem is how to evaluate $P(O|\lambda_{ad})$, the probability that the sequence was generated by HMM λ_{ad}. This probability is calculated by using the forward or backward algorithm [9].

5. Experimental Results and Discussion

5.1 Database

Experiments are performed on image sequences that have 320×240 pixel resolution and 30 frames per second. We use the Korea University (KU) gesture database[5]which contains 14 representative full body actions in the daily life of 20 performers. In the database, all the performers are elderly people (both male and female) with their age ranges from 60 to 80. The database contains 3D data and 2D data. We use the 2D video data for analysis. Our training data set includes three views such as 0^0, -45^0 and $+45^0$, respectively. The testing set can be any arbitrary view. The duration of each action depends on the type of action, which has a range of $2 \sim 8$ seconds.


(a) Walking at a place

(b) Raising a right hand

(c) Bowing

(d) Sitting on a floor

(e) Running at a place

Figure 6. Typical image sequences of human actions.

To apply HMMs, the features are transformed into symbol sequences, O in the learning and recognition phases. For each frame I_t of an image sequence, the feature vectors are extracted and these vectors are assigned to symbols sets. We use vector quantization for this implementation. We use five actions for training and testing purposes, such as (1) walking at a place, (2) raising the right hand, (3) bowing, (4) sitting on the floor, and (5) running at a place. The typical images of these actions are shown in Fig. 6. We use 16 image sequences of each action for training and 16 sequences of each action for testing. For training HMM, we use all views and for testing we use any arbitrary view.

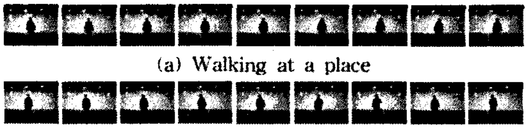
5.2 Classification experiments

For each of the five actions to be detected, we use the same topology of all HMM models, i.e. five-states fully connected models. We train discrete HMM for each action and each viewing direction with the corresponding image sequences. Classification is finally achieved by feeding a given (test) sequence to all the trained HMMs and employing a likelihood based measure to declare the action performed in the image sequence. Table 1 shows the confusion matrix of action recognition using HMM, where we use human body shape features, optical flow motion features, and combined features. Each column represents the best match for each test sequence in any arbitrary view direction. The first, second, and third value represent the recognition accuracy for shape, motion, and combined features. The average recognition rate is 87.5% which is greater than individual features. In the testing phase of experiment, we found that some sequences are misclassified, such as walking and running. These sequences are checked manually, and it is found that these image sequences are taken from front (0 degree) view. These situations are shown in Fig. 7. They may occur due to the high degree of similarity between walking and running in the image in the front view. Moreover, all the performers are old human and naturally their walking motion and the running motion are almost not so different. In the front view, the problem occurs for recognition correctly, but in the side views, it is easier to distinguish by both features.

Table 1. Confusion matrix. Action recognition using shape features, motion feature and combined features.

	walking			raise hand			bowing			running			sitting		
walking	13	12	14	1	1	0	0	0	0	2	3	2	0	0	0
raise hand	2	2	2	12	11	13	0	0	0	2	3	1	0	0	0
bowing	0	0	0	0	0	0	15	15	16	0	0	0	1	1	0
running	3	3	3	1	2	0	0	0	0	12	11	13	0	0	0
sitting	1	2	1	0	0	0	1	0	1	1	1	0	13	13	14

We also checked the recognition accuracy with some other previous researches. In the work of Ali et al. [6], they used the angles of three body components as features to recognize seven actions (walking, sitting, standing up, bending, getting up, squatting, rising) from profile views. They reported 78.8% recognition rate. Masaud et al [13] reported a recognition rate of 92.8% and use motion feature to recognize eight actions, such as walk, run, skip, march, line-walk, hop, side-walk, side-skip, respectively. Yacoob and Black [15] reported a recognition rate of 82% and recognize four actions, such as walking, line-walking, marching, and walking to kick, respectively. In [8], authors used the distribution of motion over the image space, x and y, to recognize five actions (sitting down, getting up, raising the hand, nodding, shaking hand) and obtain 66% recognition rate.


(a) Walking at a place

(b) Running at a place

Figure 7. Image sequence from front view.

In our method, it is shown that by using combined optical flow feature and human body silhouette feature gives better results. The important thing is to note that we recognize human action from any arbitrary view rather than any specific view. We only get a recognition rate lower when the person performs action in front view.

6. Conclusions and Future Research

In this research, we proposed a human action recognition method from multiple views image sequences by using human body shape features and optical flow motion features. Based on the individual feature and combined features, a set of HMMs were built for each action to represent each action from different views to enable recognizing from arbitrary views. In experiments, we compared to use only optical flow feature, the silhouette features, and combined features to build HMMs. The average recognition rate of combined features (87.5%) is higher than the rate obtained by individual features. This result showed that our algorithm is robust to variations in view and duration. Although this rate was found lower than some previous researches. But it would be mentioned that we recognize action from arbitrary views rather than any specific view. Our future work includes the interaction of multi-view learning using adaptable hidden Markov model and use complex actions.

References

[1] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," IEEE Trans. on SMC - Part C: Applications and Review, Vol. 34, No. 3, 2004, pp. 334-352.

[2] B. K. Horn and B. G. Schunck, "Determining optical flow," Artificial Intelligence, Vol. 17, 1981, pp. 82-98.

[3] M. Turk and A. Pentland, "Eigenfaces for recognition," Journal of Cognitive Neuroscience, Vol. 3, No. 1, 1991, pp. 71-80.

[4] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in a time sequential image using hidden Markov model," Proc. of IEEE Conf. on CVPR, 1992, pp. 379-385.

[5] B. -W. Hwang, S. Kim, and S.-W. Lee, "A Full-body gesture database for automatic gesture recognition," Proc. 7th IEEE Int'l Conf. on FGR, Southampton, UK, April 2006, pp. 243-248. The KU Gesture Database, http://gesturedb.korea.ac.kr.

[6] A. Ali, J. K. Aggarwal, "Segmentation and recognition of continuous human activity," Proc. of IEEE Workshop on Detection and Recognition of Events in Video, Canada, July, 2001, pp. 28-35.

[7] V. Parameswaran and R. Chellappa, "View invariants for human action recognition," Proc. of IEEE CS Conference on CVPR, Vol. 2, 2003, pp. 613-619.

[8] M. Zobl, F. Wallhoff, and G. Rigoll, "Action recognition in meeting scenarios using global motion features," Proc. of IEEE Workshop on Performance Evaluation of Tracking and Surveillance, 2003.

[9] R. Lawrence, and A. Rabiner, "Tutorial on hidden Markov models and selected applications in speech recognition," Proc. of IEEE, Vol. 77 No. 2, 1989, pp. 257-286.

[10] S. Carlsson and J. Sullivan, "Action recognition by shape matching to key frames," Proc. of IEEE CS Workshop on Models versus Exemplars in Computer Vision, Florida, USA, 2002, pp. 263-270.

[11] X. Sun, C. Chen, and B. S. Manjunath, "Probabilistic motion parameter models for activity recognition," Proc. of IEEE Int'l Conference on Pattern Recognition ,Vol. 1, pp. 443-446, Quebec City, Canada, August, 2002.

[12] J. Foster, M. Nixon, and A. Prugel-Bennett, "Automatic gait recognition using area based matrices," Pattern Recognition Letters, Vol. 24, 2003, pp. 2489-2497.

[13] O. Masoud and N. Papanikolopoulos, "Recognizing human activities," Proc. of IEEE Conference on Advanced Video and Signal Based Surveillance, Florida, USA, July 2003, pp. 157-162.

[14] C. Rao and M. Shah, "View-Invariance in action recognition," Proc. of IEEE CS Conference on CVPR, Hawaii, December, 2001, pp. 316-323.

[15] Y. Yacoob and M. J. Black, "Parameterized modeling and recognition of activities," CVIU, Vol. 73, No. 2, February 1999, pp. 232-247.