University of Wollongong

Research Online

Faculty of Informatics - Papers (Archive)

Faculty of Engineering and Information
Sciences

2008

Graphical modeling and decoding of human actions

Wanqing Li

University of Wollongong, wanqing@uow.edu.au

Zhengyou Zhang

Microsoft Research, zhang@microsoft.com

Zicheng Liu

Microsoft Research, zliu@microsoft.com

Follow this and additional works at: <https://ro.uow.edu.au/infopapers>

Part of the [Physical Sciences and Mathematics Commons](#)

Recommended Citation

Li, Wanqing; Zhang, Zhengyou; and Liu, Zicheng: Graphical modeling and decoding of human actions
2008.

<https://ro.uow.edu.au/infopapers/3170>

Research Online is the open access institutional repository for the University of Wollongong. For further information
contact the UOW Library: research-pubs@uow.edu.au

Graphical modeling and decoding of human actions

Abstract

This paper presents a graphical model for learning and recognizing human actions. Specifically, we propose to encode actions in a weighted directed graph, referred to as action graph, where nodes of the graph represent salient postures that are used to characterize the actions and shared by all actions. The weight between two nodes measures the transitional probability between the two postures. An action is encoded as one or multiple paths in the action graph. The salient postures are modeled using Gaussian mixture models (GMM). Both the salient postures and action graph are automatically learned from training samples through unsupervised clustering and expectation and maximization (EM) algorithm. Experimental results have verified the performance of the proposed model, its tolerance to noise and viewpoints and its robustness across different subjects and datasets.

Disciplines

Physical Sciences and Mathematics

Publication Details

Li, W., Zhang, Z. & Liu, Z. (2008). Graphical modeling and decoding of human actions. International Workshop on Multimedia Signal Processing (pp. 175-180). Los Alamitos, California, USA: IEEE.

Graphical Modeling and Decoding of Human Actions

Wanqing Li ^{#1}, Zhengyou Zhang ^{*2}, Zicheng Liu ^{*3}

[#] SCSSE, University of Wollongong, NSW 2522, Australia

¹ wanqing@uow.edu.au

^{*} Microsoft Research, Redmond, WA 98052, USA

² zhang@microsoft.com ³ zliu@microsoft.com

Abstract—This paper presents a graphical model for learning and recognizing human actions. Specifically, we propose to encode actions in a weighted directed graph, referred to as *action graph*, where nodes of the graph represent salient postures that are used to characterize the actions and shared by all actions. The weight between two nodes measures the transitional probability between the two postures. An action is encoded as *one or multiple paths* in the action graph. The salient postures are modeled using Gaussian Mixture Models (GMM). Both the salient postures and action graph are automatically learned from training samples through unsupervised clustering and expectation and maximization (EM) algorithm. Experimental results have verified the performance of the proposed model, its tolerance to noise and viewpoints and its robustness across different subjects and datasets.

I. INTRODUCTION

In the recent years, silhouettes have gained increasing attention in human motion analysis due to the advances in background modeling for the extraction of silhouettes, their ability to capture the spatiotemporal characteristics of human motion, and possibly lower complexity of computation. This paper is about the recognition of human motion based on sequences of silhouette images. In particular, we focus on the recognition of *human actions*, the smallest recognizable semantically meaningful motion units, such as *run*, *walk* and *jump*.

An action recognition system is desired to be independent of the subjects who perform the actions, independent of the speed at which the actions are performed, robust against noisy extraction of silhouettes. In this paper, we propose a graphical model of human actions. Specifically, we characterize actions with sequences of finite salient postures and propose to model the dynamics or kinematics of the actions using a weighted directed graph, referred to as *action graph*, and to model the salient postures with Gaussian Mixture Models (GMM). In such an action graph, nodes represent salient postures that are shared by the actions and the weight between two nodes measures the transitional probability between the two postures represented by the two nodes. This transitional probability is effectively governed by the kinematics of the human body. An action is encoded in *one or multiple paths* in the action graph. The GMM model of the salient postures provides a compact

description of the spatial distribution of the contours belonging to the same salient posture and robust matching to imperfect or noisy silhouettes.

The proposed modeling system is substantially differentiated from and possesses advantages over the previously proposed methods based on postures (or key-frames) [1], [2], [3], [4] and Hidden Markov Model (HMM) [5], [6], [7]. Firstly, our model shares postures among the actions and, hence, enables efficient learning from a small number of samples rather than modeling each action with individual HMM which often requires large number of samples to train. Secondly, we encode one action into multiple paths (or sequences of salient postures in the graph) to accommodate the variations of the action (e.g. performed by different persons or captured from different viewpoints) as opposed to one sequence of postures (or key-frames) as featured in most methods proposed so far. Thirdly, there are no specific beginning or ending postures for any action path. This allows continuous recognition of actions without segmentation. Moreover, cyclic and non-cyclic actions can be dealt with in the same way. Lastly, the model facilitates five different action decoding schemes (as described in Section III-B) that require different computing resources. From this perspective, our model can be considered as a generalization of the previous works which usually employ only one of the decoding schemes. Performance evaluation of the proposed graphical model and algorithms is carried out on a relatively large dataset currently widely used in the research community not only through the leave-one-sample-out test, but also the leave-one-subject-out and cross-dataset test (i.e. training and test data are from different datasets). The results have verified that the proposed model is able to recognize actions effectively and accurately.

The rest of the paper is organized as follows. Section II gives a brief review of previous work. Section III details the proposed graphical model of actions and the five different decoding schemes derived from the model. In Section IV, system learning algorithms are described, which include modeling of the salient postures using GMM and construction of the action graph. Experimental results and comparison of the five action decoding schemes are presented in Section V. The paper is concluded with remarks and future work in Section VI

II. RELATED WORK

A rich palette of diverse ideas has been proposed during the past few years on the problem of recognition of human actions by employing different types of visual information. A good review can be found in [8], [9], [10]. This section presents a review of the work related to silhouette based action recognition.

Study of the kinematics of human motion suggests that a human action can be divided into a sequence of postures [11], [12]. Methods proposed so far for silhouette based action recognition differs in the way that the postures are described and the dynamics of the posture sequence is modeled. In general, they fall into two categories based on how they model the dynamics of the actions: *implicit* and *explicit* models. In an implicit model, *action descriptors* are extracted from the action sequences of silhouettes such that the action recognition is turned from a temporal classification problem to a static classification one. Proposed action descriptors include moments of Motion Energy Images (MEI) and Motion-History Images (MHI) calculated from silhouettes [6], [13], GMMs to capture the distribution of the moments of the silhouettes [7] or the five extremities [3] corresponding to the arms, legs and head over the period of an action, ignoring the temporal order of the silhouettes in the action sequence, an ensemble of GMMs of category feature vectors (CFV) [14], the differential geometric surface properties [15] of the spatiotemporal volume formed by the sequence of silhouettes, and the space-time features [16] by utilizing the properties of the solution to the Poisson equation.

The implicit modeling approach has the advantages that the recognition is relatively simple and is able to handle small number of training samples. However, it usually offers weak encoding of the action dynamics and requires good temporal segmentation before the actions can be recognized. In addition, periodic or cyclic actions have to be dealt with differently [17].

On the other hand, the explicit model follows the concept that an action is composed of a sequence of postures and usually consists of two components: description of the postures and modeling of the dynamics of the postures. Divis and Tyagi [7] used moments to describe shapes of a silhouette and continuous HMM to model the dynamics. In [4], Kellokumpu et al. chose Fourier shape descriptors and classified the postures into a finite number of clusters. Discrete HMM are then used to model the dynamics of the actions where the posture clusters are considered to be the discrete symbols emitted from the hidden states. Sminchisescu et al. [18] relaxed the HMM assumption of conditional independence of observations given the actions by adopting the Conditional Random Field (CRF) model. Veerarahavan, et al. [19] proposed to use autoregressive (AR) model and autoregressive and moving average (ARMA) model to capture the kinematics of the actions. They adopted Kendall's representation of shape as shape features. Recently, Wang and Suter [17] employed Locality Preserving Projection (LPP) to learn a subspace to describe the postures and DTW and temporal Huasdorff distance to classify the actions in the

subspace. Colombo, et al. [20] proposed to find the subspace for each type of actions through Principal Component Analysis (PCA).

Lv and Nevatia [2] took the approach a step further. They modeled the dynamics using an *unweighted directed graph*, referred to as action net, where nodes in the graph represented key-postures learned from simulated actions based on the data captured from motion capture devices. The direct links indicate the allowed transition between postures. Each action is represented by *one path* in the action graph. Given an input sequence of silhouettes, the likelihood of each frame belonging to every postures is computed and the input is recognized as the action which gives the maximum accumulated likelihood along the path of the action. Similar to the implicit model, most proposed explicit modeling approaches mentioned above also require segmentation of the actions from the input sequence of silhouettes before an action can be recognized. In addition, the dynamics of the actions are modeled individually and separately (i.e. no connection among actions), such as the conventional HMM based approach. As a result, they often require a large number of training samples, which can be costly and tedious to obtain.

III. GRAPHICAL MODELING AND DECODING OF ACTIONS

Let $X = \{x_1, x_2, \dots, x_n\}$ be a sequence of n silhouettes, $\Omega = \{\omega_1, \omega_2, \dots, \omega_M\}$ be the set of M salient postures that constitute actions. The corresponding posture sequence derived from X is denoted as $S = \{s_1, s_2, \dots, s_n\}$, where $s_t \in \Omega, t = 1, 2, \dots, n$. Assume the $\Psi = \{\psi_1, \psi_2, \dots, \psi_L\}$ denotes a set of L actions and X is generated from one of the L actions. The recognition of the most likely action that generates the observation of X can be formulated as

$$\begin{aligned} \phi^* &= \arg \max_{\phi \in \Psi, S \subset \Omega} p(X, S, \phi) \\ &\propto \arg \max_{\phi \in \Psi, S \subset \Omega} p(\phi)p(S|\phi)p(X|S, \phi) \\ &= \arg \max_{\phi \in \Psi, S \subset \Omega} p(\phi)p(s_1, \dots, s_n|\phi)p(x_1, \dots, x_n|s_1, \dots, s_n, \phi), \end{aligned} \quad (1)$$

where $p(\phi)$ is the prior probability of action ϕ , $p(S|\phi)$ is the probability of S given action ϕ and $p(X|S, \phi)$ is the probability of X given S and ϕ .

Assume that i) x_t is statistically independent of ϕ given S , ii) x_t statistically depends only on s_t and iii) s_t is independent of the future states and only depends on its previous state s_{t-1} . Then, Eq. (1) can be written as

$$\phi^* = \arg \max_{\phi \in \Psi, S \subset \Omega} \underbrace{p(\phi)p(s_1, s_2, \dots, s_n|\phi)}_{\text{specific-knowledge}} \underbrace{\prod_{t=1}^n p(x_t|s_t)}_{\text{shared-knowledge}}, \quad (2)$$

where $p(x_t|s_t)$ is the probability for x_t to be generated from state or salient posture s_t . It is referred to as *posture or state model*. Contrary to conventional HMM, we assume the set of postures is known or can be computed from training data.

Fig. 1. An action graph for three actions with nine postures. In each graph, the number next to the highlighted links (in red) is the transitional probabilities. (a) Action Run; (b) Action Walk; (c) Action Side

A. Action graph

Eq.(2) can be represented or interpreted as a set of weighted directed graphs, G , that are built upon the set of postures.

$$G = \{\Omega, A, A_1, A_2, \dots, A_L\}, \quad (3)$$

where each posture serves as a node, $A_k = \{p(\omega_j | \omega_i, \psi_k)\}_{i,j=1:L}^{k=1:L}$ is the transitional probability matrix of the k 'th action and $A = \{p(\omega_j | \omega_i)\}_{i,j=1}^M$ is the global transitional probability matrix of all actions. We refer to G as an *Action Graph*.

In an action graph, each action is encoded in one or multiple paths. Figure 1 shows an action graph for three actions: *Run*, *Walk* and *Side*. The three actions share nine states/postures. Clearly, the three actions share postures and each action has multiple paths in the action graph. In addition, action paths in the graph are usually cyclic and, therefore, there are no specific beginning and ending postures/states for the action from the recognition point of view.

With the graphical interpretation, a system that follows the model Eq.(2) can be described by a quadruplet,

$$\Gamma = (\Omega, \Lambda, G, \Psi), \quad (4)$$

where

$$\begin{aligned} \Omega &= \{\omega_1, \omega_2, \dots, \omega_M\} \\ \Lambda &= \{p(x|\omega_1), p(x|\omega_2), \dots, p(x|\omega_M)\} \\ G &= (\Omega, A, A_1, A_2, \dots, A_L) \\ \Psi &= (\psi_1, \psi_2, \dots, \psi_L). \end{aligned} \quad (5)$$

B. Action decoding

Given a trained system $\Gamma = (\Omega, \Lambda, G, \Psi)$, The action of a sequence $X = \{x_1, x_2, \dots, x_n\}$ is generally decoded in three major steps: i) find the most likely path in the action graph, G , that generates X ; ii) compute the likelihood of each action, $\phi \in \Psi$; and, iii) decode the action as the one having the maximum likelihood and its likelihood is greater than a threshold, otherwise, the action of X is unknown. Eq.(2) offers a number of ways to find the most likely path and estimate the likelihood.

1) *Action Specific Viterbi Decoding*: The most obvious one is to search for an Action Specific Viterbi Decoding (ASVD) in the action graph and calculate the likelihood as follows,

$$L(\psi_i) = \max_{\phi \in \Psi, S \in \Omega} p(\phi) \prod_{t=1}^n p(s_t | s_{t-1}, \phi) \prod_{t=1}^n p(x_t | s_t), \quad (6)$$

where $L(\psi_i)$ is the likelihood of X belonging to action ψ_i . X is decoded as action ψ_k if the following condition is met

$$k = \arg \max_i L(\psi_i) \quad \text{if} \quad \frac{L(\psi_k)}{\sum_{i=1}^L L(\psi_i)} > TH_l, \quad (7)$$

where TH_l is a threshold.

2) *Global Viterbi Decoding*: In GVD, the most likely path is the one, $s^* = \{s_1^*, s_2^*, \dots, s_n^*\}$, that satisfies

$$s^* = \arg \max_{s_t \in \Omega} \prod_{t=1}^n p(s_t | s_{t-1}) p(x_t | s_t). \quad (8)$$

The likelihood of an action that generates s^* can be computed either using uni-gram or bi-gram model as below

$$L(\phi_i) = \arg \max_{\phi \in \Psi} p(\phi) \prod_{t=1}^n p(s_t^* | \phi) \quad \text{uni-gram}, \quad (9)$$

$$L(\phi_i) = \arg \max_{\phi \in \Psi} p(\phi) \prod_{t=1}^n p(s_t^* | s_{t-1}^*, \phi) \quad \text{bi-gram} \quad (10)$$

GVD decoding only requires about $\frac{1}{L}$ computational resources of what is required by ASVD.

3) *Maximum Likelihood Decoding (MLD)*: Both ASVD and GVD require memory to buffer previous frames for Viterbi search. A decoding method that does not require buffering can be devised by searching for the sequence of most likely states/postures rather than the most likely sequence of states (Viterbi path), i.e.

$$s^* = \arg \max_{s_t \in \Omega} \prod_{t=1}^n p(x_t | s_t). \quad (11)$$

The likelihood of an action to generate the path s^* can be calculated using either Eq.(9) or Eq.(10).

In all, we have five different decoding methods: 1) Action Specific Viterbi Decoding (ASVD); 2) Uni-gram with Global Viterbi Decoding (UGVD); 3) Bi-gram with Global Viterbi Decoding (BGVD); 4) Uni-gram with Maximum Likelihood Decoding (UMLD); and 5) Bi-gram with Maximum Likelihood Decoding (BMLD)

IV. SYSTEM LEARNING

Learning a system Γ from training samples involves the estimation of the posture models, Λ , and construction of the action graph, G .

Fig. 2. Feature extraction. (a) A typical silhouette; (d) Normalized and resampled points of the contour; (c) The ellipse fitted to the contour and gravity center

1) *Posture models*: A posture represents a set of similar poses. The set of postures Ω are obtained by clustering the sample silhouettes into M clusters and fitting a GMM to each cluster. Considering the temporal nature of the human motion, the clustering is based on the joint shape and motion dissimilarity between two poses, rather than shape or motion alone as used in most extant work [2], [1].

For the sake of scale invariance and noise tolerance, we choose a set of points on the silhouette contour after scale normalization as the shape descriptor. As shown in Figure 2(b), the contour of a silhouette is first normalized and then resampled to a small number of points with two purposes: noise and computation reduction.

Let $f_{sp} = \{x_1, x_2, \dots, x_b\}$ and $f'_{sp} = \{x'_1, x'_2, \dots, x'_b\}$ be the two shapes described by a set of b points on the contours respectively, the dissimilarity, d_{sp} is calculated from the Hausdorff distance between the two sets of points.

For motion dissimilarity, the change of the orientation of the entire body and the local motion of its gravity center are selected as motion features. The orientation of the body is estimated by fitting an ellipse into the silhouette shape as shown in Figure 2(c).

Let $f_m = (\delta x, \delta y, \delta \theta)$ and $f'_m = (\delta x', \delta y', \delta \theta')$ be the motion feature vector of silhouette x and x' respectively, where $(\delta x, \delta y)$ is the locomotion of the gravity center and $\delta \theta$ is the change of orientation. The dissimilarity, d_{mt} is calculated based on the correlation of f_m and f'_m .

We define the overall dissimilarity of two silhouettes as the product of their motion and shape dissimilarity. Let $D = [d_{ij}]_{i,j=1}^J$ be the dissimilarity matrix of all pairs of the J training silhouettes, where D is a $J \times J$ symmetric matrix. The J silhouettes are then clustered into M clusters by employing the traditional Non-Euclidean Relational Fuzzy (NERF) C-Means [21].

After the clustering, a Gaussian Mixture Model (GMM) is fitted using expectation and maximization (EM) algorithm to the shape component of a cluster to represent the spatial distribution of the contours of the silhouettes belonging to the same posture cluster, and one Gaussian is fitted to its motion component to obtain a compact representation of the posture models.

Let

$$p_{sp}(y_{sp}|s) = \sum_{k=1}^C \pi_{k,s} N(y_{sp}; \mu_{k,s}; \Sigma_{k,s}) \quad (12)$$

$$p_{mt}(y_{mt}|s) = N(y_{mt}; \mu_{mt,s}; \Sigma_{mt,s}) \quad (13)$$

be respectively the GMM with C components for shape and Gaussian for motion, where s represents the s salient posture/state or cluster of the silhouettes, $N(\cdot)$ is a Gaussian function; y_{mt} represents the motion feature vector; $\mu_{mt,s}$ is the mean motion vector for salient posture s ; $\Sigma_{mt,s}$ is a 3×3 matrix denoting the covariance of the motion features; y_{sp} represents the 2D coordinates of a point on the contours of silhouettes; $\mu_{k,s}$ is the center of the k 'th Gaussian for state s ; $\Sigma_{k,s}$ is a 2×2 covariance matrix; $\pi_{k,s}$ is the mixture proportion, $\sum_{k=1}^C \pi_{k,s} = 1$.

The posture model can then be defined as

$$p(x|s) = \prod_{i=1}^b p_{sp}(y_{sp}^i|s) p_{mt}(y_{mt}|s) \quad (14)$$

where x is a silhouette, y_{mt} and y_{sp}^i represent respectively the motion feature and the i 'th point on the resampled contour of x .

2) *Action Graph*: The action graph is built by linking the postures with their transitional probabilities. We estimate the action specific and global transitional probability matrices, $\{A_i\}_{i=1}^L$ and A , from the training samples given the statistical independence assumptions introduced in Section III and the posture models.

V. EXPERIMENTAL RESULTS

A. Datasets

We evaluated our model on the most widely used dataset created by Blank et al. [22]. The dataset contains 93 low resolution video (188×144 , 25 fps) sequences for 10 actions. These 10 actions are *run*, *walk*, *wave with one hand*, *wave with two hands*, *galloping sideways*, *jumping-in-place*, *jumping*, *jumping jack*, *bend* and *skip*. Nine subjects played each action once (with an exception that one subject played 3 actions twice). Silhouettes were obtained using simple background subtraction in color space. Global motion was removed by fitting quadratic function to the trajectory of the gravity centers. This dataset is currently the most realistic and challenging one publicly available compared to those employed in other papers (e.g. [23]). Some silhouettes are noisy as shown in Figure 3. Action *walk* and *jumping-in-place* appears very similar to action *galloping sideways* and *jumping* respectively when the global motion is removed from the silhouettes.

B. Experimental Results

1) *Setup*: As adopted in most previous works [22], [16], [17], [24] using the same dataset, we conducted leave-one-sample-out test to verify the overall performance of the proposed model. To evaluate its robustness against various factors including the dependence on subjects, viewpoints, action speed

Fig. 3. Examples of noisy silhouettes

and styles and video capturing environment, we also conducted the following experiments:

- leave-one-subject-out test;
- robust test against viewpoints and action styles for action *walk* using the sequences designed by Blank et al. [22], [16]; and,
- cross-dataset test. In this test, we trained an action graph using Blank’s dataset and employed the action graph to recognize 68 sequences of actions *walk* and *run* extracted from the video sequences made available by Laptev et al. [25].

In all experiments, silhouette contours were sampled to 64 points after normalization and GMMs with 32 spherical Gaussians were fitted to the shape of the contours. The following summarizes the experimental results.

2) *Leave-one-sample-out Test*: In the leave-one-sample-out test, each sample was taken as the test sample and the residual samples were used as training samples to train the action graph. Recognition rate was calculated over all samples in the dataset. Figure 4(a) shows the recognition rates of the five decoding schemes vs. number of postures, M . As expected, the two bi-gram decoding schemes (BMLD & BGVD) outperformed the two uni-gram schemes (UMLD & UGVD). The ASVD consistently outperformed both uni-gram and bi-gram decoding schemes for all M . Notice that the recognition rates of all decoding methods increase as the number of postures increases. when $M \geq 20$, the recognition rates are all above 90%. When $M = 45$, the recognition rates of BMLD, BGVD and ASVD have reached 97.8%, which are comparable to the best accuracies obtained in [22], [16], [17] and better than the accuracy (92.6%) achieved in [24].

3) *Leave-one-subject-out Test*: In the leave-one-subject-out test, the training dataset contained the samples of other actions performed by the same subject. This certainly helps the action graph to capture the styles of the postures performed by the subject and therefore benefits recognition. In the leave-one-subject-out test, we purposely took all samples performed by the same subject as the test samples and the samples performed by other subjects as the training samples. In other words, the trained action graph did not have any knowledge about the test subject. In addition, there were less number of training samples compared to the leave-one-sample-out test. Figure 4(b) shows the recognition rates of the five decoding schemes vs. number of postures, M . The curves demonstrate similar patterns to those of the leave-one-sample-out test. BMLD, BGVD and ASVD has achieved recognition accuracies of 97.8% at $M = 60$.

Since both leave-one-sample-out test and leave-one-subject-

(a) Leave-one-sample-out test

(b) Leave-one-subject-out test

Fig. 4. Recognition rates vs number of postures. (a) Leave-one-sample-out test; (b) Leave-one-subject-out test

Fig. 5. Sample silhouettes of action *moonwalk*

out test have shown that bi-gram and action specific Viterbi decoding schemes are preferred to the uni-gram decoding schemes, we excluded the uni-gram decoding schemes from the following experiments.

4) *Robustness Test*: Together with the action dataset, Blank et al [22] also supplied additional 20 samples of the action *walk* captured from 10 different viewpoints (0 degree to 81 degree relative to the image plan with steps of 9 degrees) and 10 different styles (normal, walking in a skirt, carrying briefcase, limping man, occluded Legs, knees Up, walking with a dog, sleepwalking, swinging a bag, occluded by a "pole"). We trained an action graph with 30 postures using the 93 samples for the 10 actions (none of the 20 *walk* samples were included in the training data), BMLD, BGVD and ASVD all recognized most samples and only failed to recognize the action in the cases of 72 and 81 degree viewpoints and "moonwalk" (walking with arms being raised to horizontal position). As shown in Figure 5, it is probably not unreasonable to consider the "moonwalk" as another type of action. Noticed that "Occluded by a pole" was excluded in the test since the silhouettes in this case consist of disconnected regions and our method assumes the silhouette is a connected region.

5) *Cross-dataset Test*: We further evaluated the robustness of the proposed model by conducting a cross-dataset test. In this test, we trained an action graph using Blank’s dataset and employed it to recognize the action samples from a different dataset. We chose the dataset (video sequences) made

available by Laptev [25]. The dataset comes as uncompressed video sequences with spatial resolution of 160×120 pixels and comprises six actions (*walking*, *jogging*, *running*, *boxing*, *hand waving* and *hand clapping*) performed by 25 subjects. Each subject performed each action in 4 different scenarios: 0 degree viewpoint, scale variations (from different viewpoints with the subject gradually approaching to or departing from the camera), different clothes (e.g. big pullovers or trench coats) and lighting variations. Two of the six actions, *walking* and *running*, overlap with the actions of Blank's dataset. We implemented a simple median filtering based background modeling to extract the silhouettes. Since many sequences have severe jitter, the median filter failed to extract the silhouettes. Nevertheless, we managed to extract 36 samples of action *walk* and 32 samples of action *run*. These samples were performed by 6 different subjects. BMLD, BGVD and ASVD achieved respectively 100%, 97.1% and 95.6% recognition accuracy when the action graph was trained with $M=60$.

VI. CONCLUSION AND FUTURE WORK

Recognition of human actions are still in its infancy compared to other intensively studied topics like human detection and tracking. This paper has presented a graphical model of human actions and GMM modeling of postures. Experiments have verified that the proposed model is robust against the subjects who perform the actions, tolerant to noisy silhouettes and, to certain degree, viewpoints and action styles. The model is easy to train with small number of samples due to the sharing of the postures amongst the actions. It is found that there is no significant difference in performance between the decoding scheme BMLD and BGVD. ASVD can outperform BMLD and BGVD when there are sufficient training samples, but the gain in the performance is at the expense of more computational complexity with less flexibility for continuous decoding of actions.

Our experiments have demonstrated that on average only about 3 to 5 postures per action were required to model the actions in the dataset. The average number of postures per action indicates the average length of the action paths in the graph. It is also noticed that an action graph of 30 postures that encodes the 10 actions has sparse global and action specific transitional probability matrices. In other words, many paths in the graph have not been utilized. This leaves much room for the action graph to be expanded with new actions. For an action graph with M postures that encodes L actions, there are on average $M^{\frac{M}{L}}$ paths with $\frac{M}{L}$ postures. For instance, there are about $30^3 = 27000$ paths with 3 postures in an action graph of $M = 30$ and $L = 10$, offering large capacity to encode a large number of actions and their variations. Our intention is to further evaluate the proposed model on a larger dataset and make the model expandable so that new actions can be learned and incorporated into a trained model without compromising the recognition of previously learned actions.

ACKNOWLEDGMENT

The authors would like to thank Dr Cha Zhang for many inspiring discussions during the course of the project. Wanqing Li would like to thank Communication and Collaboration Systems Group for offering the support and opportunities to develop the work at MSR, Redmond.

REFERENCES

- [1] Q. Wei, M. Hu, X. Zhang, and G. Luo, "Dominant sets-based action recognition using image sequence matching," in *Proc. ICIP*, vol. VI, 2007, pp. 133–136.
- [2] F. Lv and R. Nevatia, "Single view human action recognition using key pose matching and viterbi path searching," in *Proc. CVPR'07*, 2007.
- [3] D.-Y. Chen, H.-Y. M. Liao, and S.-W. Shih, "Humnn action recognition using 2-D spatio-temporal templates," in *Proc ICME*, 2007, pp. 667–670.
- [4] V. Kellokumpu, M. Pietikainen, and J. Heikkilä, "Human activity recognition using sequences of postures," in *Proc IAPR Conf. Machine Vision Applications*, 2005.
- [5] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human actions in time-sequential images using hidden markov model," in *Proc. CVPR92*, 1992, pp. 379–385.
- [6] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," *IEEE PAMI*, vol. 23, pp. 257–267, 2001.
- [7] J. W. Davis and A. Tyagi, "Minimal-latency human action recognition using reliable-inference," *Image and Vision Computing*, vol. 24, pp. 455–472, 2006.
- [8] D. M. Gavrila, "The visual analysis of human movement: A survey," *Computer Vision and Image Understanding*, vol. 73, no. 1, pp. 82–98, 1999.
- [9] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Trans. SMC-C*, vol. 34, no. 3, pp. 334–351, 2004.
- [10] T. B. Moeslund, A. Hilton, and V. Kruger, "A survey of advances in vision-based human motion capture and analysis," *Computer vision and Image Understanding*, vol. 104, pp. 90–126, 2006.
- [11] V. M. Zatsiorsky, *Kinematics of Human Motion*. Human Kinetics Publishers, 1997.
- [12] G. Zheng, W. Li, P. Ogunbona, L. Dong, and I. Kharitonenko, "Human motion simulation and action corpus," in *Proc of 1st International Conference on Digital Human Modeling, HCI International*, 2007.
- [13] H. Meng, N. Pears, and C. Bailey, "A human action recognition system for embedded computer vision application," in *Proc. CVPR*, 2007.
- [14] W. Lin, M.-T. Sun, P. Poovandran, and Z. Zhang, "Human activity recognition for video surveillance," *IEEE Trans. CSVT*, vol. (to appear), 2008.
- [15] A. Yilmaz and M. Shah, "Actions sketch: a novel action representation," in *Proc. CVPR*, vol. 1, 2005, pp. 984–989.
- [16] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Trans PAMI*, vol. 29, no. 12, pp. 2247–2253, 2007.
- [17] L. Wang and D. Suter, "Learning and matching of dynamic shape manifolds for human action recognition," *IEEE Trans. Image Processing*, vol. 16, pp. 1646–1661, 2007.
- [18] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas, "Conditional models for contextual human motion recognition," in *Proc ICCV*, 2005.
- [19] A. Veeraraghavan, A. R. Chowdhury, and R. Chellappa, "Role of shape and kinematics in human movement analysis," in *Proc CVPR'04*, 2004.
- [20] C. Colombo, D. Comanducci, and A. Del Bimbo, "Compact representation and probabilistic classification of human actions in videos," in *Proc IEEE AVSS'07*, 2007, pp. 342–346.
- [21] R. Hathaway and J. Bezdek, "NERF C-means: non-euclidean relational fuzzy clustering," *Pattern Recognition*, vol. 27, pp. 429–437, 1994.
- [22] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *ICCV*, 2005, pp. 1395–1402.
- [23] A. Veeraraghavan, R. Chellappa, and Roy-Chowdhury, "The function space of an activity," in *Proc CVPR'06*, 2006.
- [24] S. Ali, A. Basharat, and M. Shah, "Chaotic invariants for human action recognition," in *Proc CVPR'07*, 2007.
- [25] I. Laptev and T. Lindeberg, "Space-time interest points," in *Proc. ICCV*, 2003, pp. 432–439.