Dance Posture Recognition Using Wide-baseline Orthogonal Stereo Cameras

Feng Guo and Gang Qian
Department of Electrical Engineering and
Arts, Media and Engineering Program
Arizona State University
Tempe, AZ, 85287, USA
Email:{Feng.Guo, Gang.Qian}@asu.edu

Abstract

In this paper, a robust 3D dance posture recognition system using two cameras is proposed. A pair of wide-baseline video cameras with approximately orthogonal looking directions is used to reduce pose recognition ambiguities. Silhouettes extracted from these two views are represented using Gaussian mixture models (GMM) and used as features for recognition. Relevance vector machine (RVM) is deployed for robust pose recognition. The proposed system is trained using synthesized silhouettes created using animation software and motion capture data. The experimental results on synthetic and real images illustrate that the proposed approach can recognize 3D postures effectively. In addition, the system is easy to set up without any need of precise camera calibration.

1 Introduction

Movement and body pose-based interactivity has attracted tremendous attention in performing arts [1, 2]. For example, in interactive dance, movers can interact with a central control system through local and global body movements to control lights, trigger various background music as well as visual effects. In [2], an interactive dance system is implemented, which reads a set of predefined static 3D postures (i.e. body shapes) and informs audio and visual feedback engines with the pose recognition results so they can react accordingly. To obtain reliable body kinematics for 3D pose recognition, a marker-based motion capture system is deployed. Although this system has been used successfully in real-life performances, the requirement of a marker-based motion capture system has made the system dissemination difficult. In this paper, we propose a video-based robust 3D dance pose recognition system, which will enable the interactivity between the responsive environment and the subject without having to use a marker-based system.

3D pose recognition from images has been an active research area in computer vision, human computer interaction for years, including hand gesture, head gesture and body posture. According to the number of cameras needed in a recognition system, current approaches for 3D pose recognition from images can be categorized into two clusters: methods that use single views [8, 9, 4, 12, 5, 11], and those using multiple views, e.g. [3, 7, 6]. In these single-view approaches, classifiers are trained using a set of image samples and only one view is used for pose recognition. For example, in [9], horizonal and vertical projection histograms of a silhouette are used to classify the postures into four main postures and then into one of three view-based appearances. [4] explores an example-based approach to recover 2D joint positions by matching extracted image features (shape context) with those of the examples and then directly estimates the 3D body pose from the 2D configuration. The appearance of the object(hand, body) is view-dependent. Usually, a large set of samples are needed to take care of the view variation. To accelerate the recognition, in [12], a local sensitive hashing is used to achieve fast human upper body poses estimation from a large database of frontal-view exemplar images. Algorithms trying to reduce the amount of training images have also been proposed. For example, [5] describes a Discriminant-EM(DEM) algorithm which combines the supervised and unsupervised learning paradigms. This DEM approach makes use of a small set of labeled data to classify a large set of unlabeled data and refines the cluster models using EM iteratively. In [11], variations of possible shape appearances around the registered typical appearances are trained as locally-compressed feature manifold to cover large appearance examples.

Recognition ambiguity introduced by view point variation is one of the biggest challenges for single-view approaches, especially for the recognition of whole body poses. To better address this issue, multi-view approaches

Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR'06)
0-7695-2503-2/06 $20.00 © 2006 IEEE

have been used. In [3], three cameras are used to capture the silhouettes. The body configuration can then be estimated from these silhouettes according to a large number of examples stored in a motion capture database. In [7], a neural network is trained to map a 2D silhouette to 2D positions of body joints and then applied an EM algorithm to reconstruct a 3D body pose based on 2D body configurations from multiple views. In [6], the body postures are inferred from a 3D visual-hull constructed from a set of silhouettes. A 3D shape context is introduced to describe the 3D shape and a support vector machine (SVM) is used for pose classification. One of the challenges for multi-view methods is that precise external camera calibration is usually needed for 3D reconstruction.

In this paper, we propose robust 3D dance pose recognition system using two wide-baseline video cameras with approximately orthogonal looking directions, which can significantly reduce ambiguities compared with single-view approach. Silhouettes extracted from two views are used as features in a relevance vector machine (RVM) for pose recognition. An overview of the proposed approach is given in Figure 1. A list of features of the proposed system is as follows: (a) robust to viewpoint-variation; (b) no precise camera calibration is needed; and (c) silhouettes presented using Gaussian mixture model (GMM) and silhouette similarity measured by Kullback-Leibler divergence (KLD).

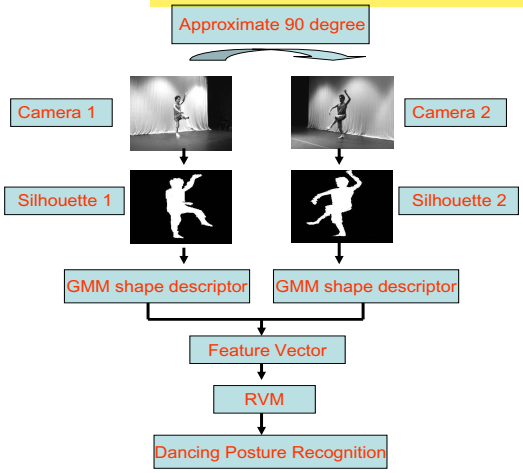

Figure 1. An overview of the proposed posture recognition system

2 Robust Silhouette Representation and Distance Measurement

2.1 Clustering Pixels into GMM

Assume that subject silhouettes can be extracted from images using background subtraction, morphological operation and image filtering. The remaining question is how to represent the silhouette robustly and efficiently. A number of silhouette representations have been proposed in the literature. For example, in [4], the shape context descriptor is used to extract the discriminative information from silhouette contour. In [15], Hu moment is applied to generate silhouette features.

Image regions are the basic building blocks in forming the visual content of an image and thus have great potentials in image representation. For example, [10] uses a multi-class statistical model of colors and shapes to obtain a 2D representation of head and hands in a wide range of viewing conditions. In [19], silhouettes of human walking is divided into seven parts and a set of moment related features on seven regions is computed. The mean of each region feature in a walking sequence is obtained as feature for gait recognition.

To increase the system's tolerance to noise present in silhouette extraction, such as shadows, a silhouette is considered to be composed of a set of 2D coherent regions and represented by a mixture of Gaussian components. GMM has been used to represent spatial point distribution in pose estimation, e.g. [17, 16]. The intuition is that the histogram occurring in an image neighborhood or window around some pixel is represented as a distribution. The Gaussian mixture model assumes that the observed unlabeled data is produced by a number of N hidden point generators, and these points follow a Gaussian distribution with a particular mean and variance. Clearly, the pixels in a silhouette image do not follow the Gaussian distribution assumption. Actually a uniform distribution confined to a closed area is a much better choice. However due to its simplicity, GMM is selected here to represent silhouette images. In addition, when the number of mixture generators is large, the image is describable with a particular mean and covariance matrix. From Figure 2, we can see that a GMM can describe the spatial distribution of the silhouette pixels fairly well. GMM-based silhouette representation is robust to silhouette extraction noise when the size of the noise is much smaller than that of the true silhouette, since such noise will only slightly change the underlying pixel spatial distribution.

Learning a Gaussian mixture model is in essence an unsupervised clustering task. The expectation-maximization (EM) algorithm has been used to solve the maximum-likelihood parameter estimation problem. Initialization is very important in EM algorithm. To make the iteration converge fast, the k-means algorithm is used to get an initial clustering of the points. Each silhouette has random initialization for EM fitting. In order to make the model invariant to translation and scaling, a silhouette image is normalized by h, the pixel height of the silhouette. Down-sampling is used in preprocessing to accelerate the computation. Different values of cluster number were used and compared.

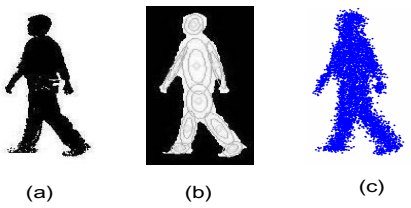

Figure 2. (left):The original silhouette. (middle): The mean and the covariance contour of Gaussian mixture components for this silhouette. (right): The reconstructed silhouette from the GMM model.

In our experiments, silhouettes were clustered into twenty mixtures. When the sum of absolute difference of the same mixture locations between two successive iterations is less than $\frac{1}{5h}$, the EM is considered converged. To make the descriptor more accurate, full covariance matrix is kept. The average computation time in Matlab code is 0.6s for each silhouette with height about 120 pixels.

2.2 Distance Measurement Using KLD

In the proposed system, the Kullback-Leibler divergence (KLD) is used to compute distance between GMMs representing two silhouettes. Similar methods have been used in content-based image retrieval and decent image matching results has been obtianed, e.g. [14].

Given two distributions p_1 and p_2, the KLD between p_1 and p_2 is:

$$D(p_1||p_2) = \int p_1(x) log \frac{p_1(x)}{p_2(x)} dx \qquad (1)$$

The KLD is not symmetric, which means that $D(p_1||p_2)$ is in general different from $D(p_2||p_1)$. The symmetric version of the KLD is:

$$d(p_1, p_2) = \frac{1}{2}[D(p_1||p_2) + D(p_2||p_1)] \qquad (2)$$

In the case of GMM, the computation of the KLD is not direct. In fact there is no analytical way to evaluate (2) and we have to resort to Monte Carlo method. The KLD between p_1 and p_2 is approximated as:

$$D(p_1||p_2) = \frac{1}{N} \sum_{i=1}^{N} log \frac{p_1(x_i)}{p_2(x_i)} \qquad (3)$$

where x_i can be either the silhouette pixels or the samples from p_1.

Figure 3 is the KLD distance map of the GMM descriptors extracted from silhouettes of 250 side-view walking images, spanning three and half gait cycles (about seven steps). The distance values are normalized by the maximum value of the distance. Dark pixels indicate small distances. A clear periodicity is present in this distance matrix, which is caused by the *half-cycle* ambiguity in side-view walking. This also indicates that the GMM descriptors captured the essential information for the similar silhouettes.

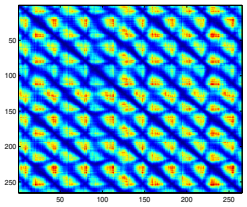

Figure 3. The silhouette distance matrix of 250 side view walking silhouettes using KLD and GMM.

Assume that at time t, two silhouettes $S_{t,1}$ and $S_{t,2}$, are obtained synchronically. Each one has the corresponding GMM representation, namely, $GMM_{t,1}$ and $GMM_{t,2}$, respectively. A combined feature vector can be defined as $F_t = (GMM_{t,1}, GMM_{t,2})$. The distance measurement of two feature vectors F_t and F'_t is the sum of the two KLDs:

$$D(t, t') = d(GMM_{t,1}, GMM_{t',1}) + d(GMM_{t,2}, GMM_{t',2}) \qquad (4)$$

3 Posture Recognition Using RVM

Given distance measurement of the silhouette images, posture recognition is a binary classification problem. A relevance vector machine (RVM) is used to train and recognize poses using the set of information provided by the GMM descriptors. The main advantage of using RVM is its ability to compress the information contained in the training set since only relevance vectors (RVs) are required for the classification. This will reduce the computation cost needed in pose recognition greatly.

Pioneered by Tipping [18], the RVM learning approach is based on Bayesian estimation theory, which can be applied to both classification and regression problems. It yields nearly identical performance to, if not better than, that of SVM in several benchmark studies [18]. Compared with SVM, RVM is found to be advantageous on several aspects [18], including: 1) the RVM decision function can be much sparser than the SVM classifier, i.e., the number of RVs can be much smaller than that of support vectors

(SVs), and 2) RVM does not need the tuning of a regularization parameter 'C' as SVM does during the training phase. This generally entails a cross-validation procedure, which is costly both of data and computation.

Consider binary classifier training. Let vector x denote a posture pattern to be classified, scalar $y, y \in \{\pm 1\}$ its class label, and $\{(x_i, y_i), i = 1, 2, \ldots, N\}$ a given set of training examples with known class labels. The problem being addressed here is the definition of a decision function that can make accurate classification of unseen values of x.

For an input vector , an RVM classifier models the probability distribution of its class label $y \in \{+1, -1\}$ using logistic sigmoid link function:

$$p(y = 1|x) = \frac{1}{1 + exp\{-f_{RVM}(x)\}} \quad (5)$$

where $f_{RVM}(x)$ is the classifier function, given by $f_{RVM}(x) = \sum_{i=1}^{N} w_i K(x, x_i)$. Here $K(.,.)$ is the kernel function , w_i are the weights of the kernel and x_i are the training samples.

The weights w_i are determined using Bayesian estimation. Toward this end, the hyper-parameters α_i are introduced to control the weights w_i. This is expressed in a Bayesian context via a Gaussian prior over W, pulling it towards zero:

$$p(W|\alpha) = \prod_{i=0}^{N} N(w_i|0, \alpha_i^{-1}) \quad (6)$$

where hyper-parameters α_i is based on the Gamma distribution. This will avoid the over-fitting and generate sparse non-zero weights.

Given the hyper-parameter control, the weights w_i in (3) are obtained by maximizing the posterior distribution of the class labels given the input vectors. This is equivalent to maximizing the following objective function:

$$J(w_1, \ldots w_N) = \sum_{i=1}^{N} logp(y_i|x_i) + \sum_{i=1}^{N} logp(w_i|\alpha_i) \quad (7)$$

where the first summation term corresponds to the likelihood of the class labels, and the second term corresponds to the prior on the parameters, in which denotes the maximum a posteriori estimate of the hyper-parameter. In the resulting solution, only those samples associated with nonzero coefficients, called relevance vectors (RVs), will contribute to the decision function.

In (3), a kernel function is used to form expansion basis functions for $f_{RVM}(x)$. Kernel function $K(x_i, x_j)$ measures the similarity between two points x_i and x_j in the space X. A popular example is the Gaussian kernel. In the proposed system, input data x are GMMs representing silhouettes, a Kullback-Leibler divergence (KLD) is used to compute $K(x_i, x_j)$. The KLD kernel in probability space is analogous to the Gaussian kernel in Euclidean space. It is non-Euclidean measures of similarity between two probability distributions. The KLD kernel has been shown to achieve very good results in the domains of object recognition [13]. The KLD kernel is defined as follows:

$$K(x_i, x_j) \Longrightarrow K(p(x_i|\theta_i), p(x_j|\theta_j)) \Longrightarrow exp\{-\frac{D(p_i, p_j)}{\beta}\} \quad (8)$$

where β is variance estimated from the distance of all training GMM distances. In our experiments, $\beta = 3$. Similar results can be obtained with β chosen as $1, 2$ and 5.

The original RVM is derived and experimented on two-class classification problems. While mentioning an extension to multi-class, the original formulation essentially treats the multi-class problems as a series of n one-vs-the-rest binary classification problems. This would translate into training n binary classifiers independently. The output of the classifier with the maximum probability is labeled as the final classification result. If none of the classifiers can provide output larger than a prechosen threshold, the input will be classified to be not a pose.

4 Experimental Results

Shown in Figure 4, 20 postures were used to train and test the proposed system. These postures were used to facilitate a story telling dance piece choreographed by Bill T. Jones.

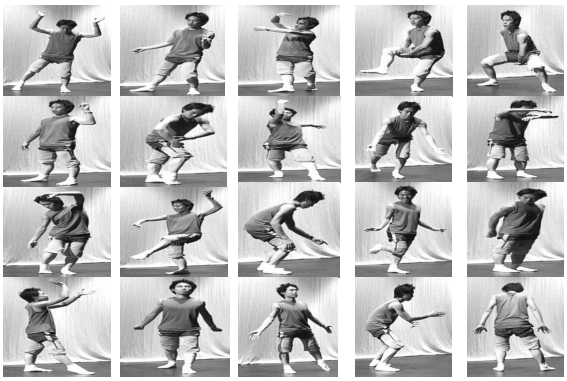

Figure 4. 20 dance postures. The posture label increases by row from left upper corner to right down corner.

The training data set consists of 207 3D shapes obtained from 3D motion capture data performed by Bill. It includes 20 posture shapes with some acceptable variations and some *trick-poses* that are similar to the true postures in some aspects but not acceptable as true poses. Each shape is ren-

dered using animation software from 16 viewpoints uniformly sampled on the sphere equator centered at the hip of the body. For a pair of orthogonal cameras, we consider counterclockwise direction to label the camera order so that the elements in input vector will have the same order with the training data. Hence the total number of input vectors is 3312. A pose is determined by the joint angles and two-degree torso orientation angles. A pose is considered unchanged if the subject turns about the axis perpendicular to the ground plane.

Regarding the learnt RVM classifiers, the number of RVs (produced during training) is found to be 298 for 20 classifiers, which is about 9% of the total training samples. The proposed classification algorithm was tested using both real videos and synthetic data. We manually established ground truth for each pair images by labeling it with the corresponding posture shape. The 3D body model used in testing silhouette generation was the same as the one used in the training.

4.1 Performance Analysis

The system performance is evaluated using recognition rate and false alarm rate. Assume that in the testing, there are N_p frames of true poses and among them, n_r frames are correctly recognized. The recognition rate is thus computed as $\frac{n_r}{N_p}$. Assume that there are N_n frames of non-poses and n_f be the number of frames that are incorrectly recognized as some of the poses. The false alarm rate is computed as $\frac{n_f}{N_p \times 19 + N_n \times 20}$.

4.1.1 Sensitivity to Torso Orientation Variations

This experiment measured the average performance of our approach with respect to torso orientation changes. In the generation of testing silhouettes, everything is the same as the training data except torso orientations. Totaly 1656 testing silhouette pairs were generated and used to test the system. The current recognition rate is 98.4% and the false alarm rate is 0.29%.

4.1.2 Sensitivity to Camera Calibration Errors

The second experiment tested the sensitivity of the system to the accuracy of angles between the camera looking directions. The generation of testing silhouettes is identical to that in the previous experiment, except camera looking direction angles were set to several values, between 75 and 105 degrees. 1656 pairs of images were created and used to test the system. The recognition rate is 94.6% and the false alarm rate is 0.37%. For different angle values, the error rates are shown in Table 1. From Table 1, we can see that as the viewing-direction angles is more close to orthogonal, the error becomes smaller.

Table 1. Error rates for different camera relative angles.

Degree	75&105	80&100	85&95	87.5&92.5
Recognition	97.8%	98.3%	99.2%	99.2%
False alarm	0.12%	0.09%	0.09%	0.08%

4.1.3 Discriminative Analysis

The third experiment measured the ability of the system to distinguish trick-poses from the true ones. The camera looking directions are orthogonal to each other. The motion capture data used to generate testing silhouettes included capture of a walking sequence, which is very different from the targeted posture and performed by another subject, and trick-poses, which are non-poses but close to the true ones, performed by the same subject as the training data. These non-poses were not used in training. Some of them are shown in the second row in Figure 5. 90 images from the walking sequence and 632 images of trick-poses. The false alarm rates for walking sequence and the trick-poses are 0 and 0.41%, respectively.

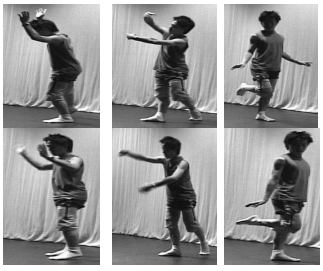

Figure 5. The first row show the true postures. The second row show their corresponding trick-postures.

4.2 Experiments Using Real Videos

Real video images were captured using two synchronized video cameras. Cameras are located at roughly mid-body height and their angular displacement is approximately 90 degrees. The intrinsic camera calibration is unknown.

To obtain the silhouettes from the image, we use background subtraction in HSV space. The real video data were captured when a dancer was performing the poses, who is different from the dancer (Bill), whose data is used in system training. It means that the body shape and the clothes in test images are different from the training. Some images from one camera are shown in Figure 4. The testing

data includes both true poses and non-poses. The torso orientation with cameras varies randomly from trial to trial. Some of the torso orientations are not included in training data. In total, 120 pairs of input images are tested. The recognition rate is 81.9% and the false alarm rate is 0.36%. The relatively large recognition error was caused mainly by body shape and pose execution differences between the two dancers. This will cause sometimes large differences between the testing and training silhouettes of the same pose. The computation time is 2.21s/frame in average using a PC with Pentium M 1.86GHz CPU with nonoptimized Matlab implementation.

5 Conclusions

A robust posture recognition system using two wide-baseline orthogonal video cameras is proposed in this paper. By deploying two cameras, ambiguities in posture recognition can be significantly reduced. GMM and KLD are used to represent posture silhouettes extracted from two views and to measure silhouettes distances, respectively. An RVM-based recognition system is trained and tested using both synthetic and real images and satisfactory results have been obtained. Experimental results also show that the proposed system is robust to small variations in system set-up, and there is no need to precise either internal or external camera calibrations.

6 Acknowledgements

The authors thank Bill T. Jones for creating the poses and Siew Wong for providing the data using the system training and testing. This paper is based upon work partly supported by National Science Foundation under Grant CISE-RI No. 0403428.[1]

References

[1] A.Camurri, S.Hashimoto, M.Ricchetti, A.Ricci, K.Suzuki, R.Trocca, and G.Volpe, "EyesWeb: Toward Gesture and Affect Recognition in Interactive Dance and Music Systems," *Computer Music Journal*. 24(1), 57-69, 2000

[2] G.Qian, F.Guo, T.Ingalls, L.Olson, J.James and T.Rikakis, "A Gesture-Driven Multimodal Interactive Dance System," *ICME*, 2004

[3] L.Ren, G.Shakhnarovich, J.Hodgins, H.Pfister and P.Viola,"Learning Silhouette Features for Control of Human Motion,"*ACM Trans.on Graphics*,vol 24, No.4, 2005

[4] G.Mori,J.Malik,"Estimating Human Body Configurations Using Shape Context,"*ECCV*,2002

[5] Y.Wu, T.S.Huang,"View-independent Recognition of Hand Postures," *CVPR*, 2000.

[6] I.Cohen, H.Li, "Inference of Human Postures by Classification of 3D Human Body Shape,"*FGR*,2003

[7] R.Rosales, M.Siddiqui,J.Alon and S.Sclaroff, "Estimation 3D Body Pose Using Uncalibrated Cameras,"*CVPR* 2001

[8] Y.Cui,L.Swets and J.Weng, "Learning-Based Hand Sign Recognition Using SHOSLIF-M,"*FGR*, 1995

[9] I. Haritaoglu, D.Harwood and L.S.Davis, "Ghost: A Human Body Part Labeling System Using Silhouettes,"*ICPR*, 1998

[10] C.Wren,A.Azarbayejani,T. Darrell and A.Pentland, "Pfinder: Real-Time Tracking of the Human Body," *PAMI*,1997

[11] A.Imai, N.Shimada, Y.Shirai, "3-D Hand Posture Recognition by Training Contour Variation,"*FGR*,2004

[12] G.Shakhnarovich,P.Viola,T.Darrell,"Fast Pose Estimation with Parameter-Sensitive Hashing," *ICCV*, 2003

[13] N. Vasconcelos, P. Ho, and P. Moreno. "The Kullback-Leibler Kernel as a Framework for Discriminant and Localized Representations for Visual Recognition," In *ECCV*, 2004.

[14] J.Goldberger, S.Gordon and H.Greenspan,"From Image Gaussian Mixture Models to Categories," *ECCV*, 2002.

[15] R.Rosales and S. Sclaroff, "Learning Body Pose Via Specialized Maps," *NIPS* 2002.

[16] E.A.Hunter, P.H.Kelly and R.C.Jain, "Estimated of Articulated Motion Using Kinematically Constrained Mixture Densities," *IEEE Proceedings Nonrigid and Articulated Motion Workshop*,1997.

[17] S. Cheng and M.Trivedi,"Hand Pose Estimation Using Expectation-Constrained-Maximization From Voxel Data," *Technical Report,CVRR Lab*, Nov,2004.

[18] M. Tipping. "Sparse Bayesian learning and the Relevance Vector Machine," *Journal of Machine Learning Reserach*, 2001.

[19] L. Lee, W. E. L Grimson, "Gait Analysis for Recognition and Classification,"*FGR*, 2002.

[1] Any opinions, findings and conclusions or recomendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF).