HUMAN MOVEMENT RECOGNITION USING FUZZY CLUSTERING AND DISCRIMINANT ANALYSIS

Nikolaos Gkalelis [#†], *Anastasios Tefas* [†] *and Ioannis Pitas* [#†]

[#] Informatics and Telematics Institute, CERTH, Greece
[†] Department of Informatics, Aristotle University of Thessaloniki, Greece
{galelis,tefas,pitas}@aiia.csd.auth.gr

ABSTRACT

In this paper a novel method for human movement representation and recognition is proposed. A movement is regarded as a sequence of basic movement patterns, the so-called dynemes. Initially, the fuzzy c-mean (FCM) algorithm is used to identify the dynemes in the input space, and then principal component analysis plus linear discriminant analysis (PCA plus LDA) is employed to project the postures of a movement to the identified dynemes. In this space, the posture representations of the movement are combined to represent the movement in terms of its comprising dynemes. This representation allows for efficient Mahalanobis or cosine-based nearest centroid classification of variable length movements.

1. INTRODUCTION

Human motion encompasses several levels of complexity. In this context, [4], proposed a three level taxonomy. In the lower level, a dyneme, is described as the most constructive unit of motion, while one level above, movement, is conceived as a sequence of dynemes, with clearly defined temporal boundaries as well as clear conceptual interpretation, e.g., one period of walk. In this paper, we adopt the above taxonomy, and in particular, we identify the dynemes in the input space by FCM clustering [1], and then express the comprising postures of a movement in terms of these dynemes using PCA plus LDA [5]. In the projection space the postures are combined to represent a movent with a single movement vector and allow simple Mahalanobis- or cosine-based nearest-centroid movement classification.

Different movements comprise a different number of postures. Moreover, the same movement is executed differently when performed from different people, or even from the same person more than one times. Most researchers in the field represent a movement as a manifold in some space and exploit an expensive similarity metric, e.g., Hausdorff distance, in order to compare movements of different length and account for internal speed variations. In contrary, with the proposed method, a movement is represented as a single vector in terms of the comprising dynemes, which implicitly takes into account internal speed variations, and allows for efficient comparison of variable length movements. This characteristic makes the method attractive for real time movement recognition applications.

2. PRIOR WORK

There is a bulk of human motion analysis literature. A general review can be found in [3]. An important question in human movement recognition is what kind of information should be exploited in order to represent a movement. Methods in the literature mostly exploit the dynamic local or global motion information within a sequence of posture images to represent motion. There have been also proposed methods that combine both local and global motion cues as well as methods that exploit static posture information alone.

Local motion information is derived by observing the spatial variation of points in the human body over time. Point correspondences are acquired either by feature tracking or optical flow, e.g., [4, 10]. Then position and velocity of these points is used to represent the motion in the input space.

Global motion refers to the shape configurations that the human body receives through the course of a movement, without considering any point correspondences between the images. Consequently, a movement is represented by a sequence of posture images extracted from the original video, e.g., [6–9]. In this paper, we represent a movement with a sequence of silhouettes, i.e., global motion information is exploited.

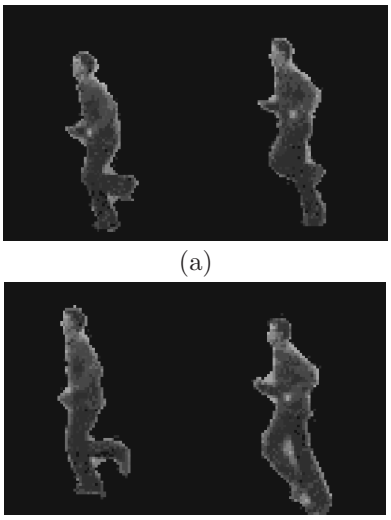
From the classification point of view, most methods mainly fall within two categories, template matching and statistical techniques. In [6], a subspace technique is used to represent a movement manifold in the feature space. A novel manifold is classified with its nearest neighbor, using median Hausdorff distance or normalized spatiotemporal correlation. In [7], a movement is represented by a sequence of space-time shape features using the Poison equation, and a novel space-time template is recognized using the nearest neighbor classifier. A clustering, Dominant Sets-based method is proposed in [8], to represent a movement with a sequence of frequency vectors, and classify a novel movement in real time.

In [4], tracking information is exploited to form motion vectors for each frame and a number of HMMs are trained to recognize a variety of movements. In [10], motion information from major body parts provided from tracking is used to form feature vectors, and an exhaustive Mahalanobis-based majority voting criterion is used to classify a novel movement.

3. PROPOSED METHOD

A movement is represented as a discrete spatiotemporal sequence of posture images. The posture image is scanned column-wise to form the so-called posture vector $\mathbf{x} \in \Re^F$ and, thus, represent the movement with a spatiotemporal trajectory in the input space $\{\mathbf{x}_\ell\}_{\ell=1:L}$.

Movement classes highly overlap, and therefore do not express the actual structure of the input space $\Re^F$. An example of such overlap is shown in Fig. 1, where a naive observer may mistakenly perceive the sequence of "skip" postures with "run". One way to counteract this problem is to represent movements as manifolds in some space. Depending on the nature of the embedded space, linear or non-linear subspace techniques can be used to learn the manifolds, e.g., as it is done in [6]. The disadvantage of these approaches is that matching manifolds usually requires an expensive similarity metric, e.g., Hausdorff distance [6–8], in order to account for time shifts and internal speed variations when comparing two movements.



(a)

(b)

Figure 1: ***(a)*** *Two postures of "skip" which can be easily confused with postures of "run".* ***(b)*** *Two postures of "run".*

In this paper, in order to recognize K different movement classes, we assume that there are C posture classes in the input space, where $C > K$. The posture classes are identified by unsupervised clustering and represented by their centroid, the so-called dyneme posture (for instance, three dyneme postures can be seen in Fig. 2). Although the C dyneme postures may belong to more than one movement classes, when combined appropriately can uniquely characterize the K different movements. Based on this assumption, we apply the FCM algorithm to discover the dynemes, and then PCA plus LDA to project the individual postures to the identified dyneme classes. In this space movement parts can be represented by the arithmetic mean of the comprising postures and uniquely characterize each movement type.

3.1 FCM to discover input space structure

We assume that the number of dynemes in the input space is $C > K$. Considering unlabelled posture data $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, we apply the FCM algorithm [1] to discover the intrinsic structure of the input space. The FCM algorithm is based on the minimization of the following objective function:

$$\mathbf{J}_{FCM}(\mathbf{\Phi}, \mathbf{V}) = \sum_{c=1}^{C} \sum_{\imath=1}^{N} (\phi_{c,\imath})^m (\| \mathbf{x}_\imath - \mathbf{v}_c \|_2)^2 \,, \quad (1)$$

where, N, C are the number of samples and centroids respectively, $\mathbf{x}_\imath \in \Re^F$ is the $\imath$-th sample in the training data set, $\mathbf{V} = [v_{\jmath,c}] = [\mathbf{v}_1, \ldots, \mathbf{v}_C] \in \Re^{F \times C}$ is the matrix of cluster prototypes, in our case the dyneme representations, $\mathbf{\Phi} = [\phi_{c,\imath}] \in \Re^{C \times N}$ is the partition matrix with $\phi_{c,\imath} \in [0, 1]$ the degree that the $\imath$-th sample belongs to the c-th cluster, $m > 1$ is the fuzzification parameter and $\| \cdot \|_2$ is is the euclidian vector norm. The FCM criterion (1) is subjected on producing non-degenerate fuzzy partition of the training data at each iteration of the optimization, $\{\mathbf{\Phi} \in \Re^{C \times N} \mid \forall c, \imath : \sum_{c=1}^{C} \phi_{c,\imath} = 1;\ 0 < \sum_{\imath=1}^{N} \phi_{c,\imath} < N;\ 0 \leq \phi_{c,\imath} \leq 1\}$.

The computation of the cluster centers and partition matrix is carried out through iterative optimization of (1), with the update of membership matrix and cluster centers at each step given by:

$$\phi_{c,\imath} = \frac{(\| \mathbf{x}_\imath - \mathbf{v}_c \|_2)^{-2(m-1)^{-1}}}{\sum_{\jmath=1}^{N} (\| \mathbf{x}_\jmath - \mathbf{v}_c \|_2))^{-2(m-1)^{-1}}}, \quad (2)$$

$$\mathbf{v}_c = \frac{\sum_{\imath=1}^{N} \phi_{c,\imath}^m \mathbf{x}_\imath}{\sum_{\imath=1}^{N} \phi_{c,\imath}^m} \,. \quad (3)$$

The iteration is initialized with an initial estimate of matrix $\mathbf{V}$ or $\mathbf{\Phi}$ and terminates when the difference of the estimated matrix between two iterations is smaller than a specified tolerance ϵ.

3.2 Projecting to dyeneme classes with PCA plus LDA

FCM algorithm will assign posture vector $\mathbf{x}_\imath$ to a dyneme class $c = 1, \ldots, C$ with membership degree, $\phi_{c,\imath}$. Based on the clustering results we label each posture $\mathbf{x}_\imath$ according to the cluster it is assigned with the largest membership degree, and weigh it with the corresponding membership degree

$$o = \underset{c \in [1,\ldots,C]}{\operatorname{argmax}} (\boldsymbol{\phi}_{c,\imath}) \,, \quad (4)$$

$$\mathbf{z}_\imath^{(o)} = \phi_{o,\imath} \mathbf{x}_\imath \,, \quad (5)$$

to produce a set of labelled data $\{\mathbf{z}_1^{(1)}, \ldots, \mathbf{z}_{N_1}^{(1)}, \ldots, \mathbf{z}_{N_C}^{(C)}\}$, where N_c is the number of postures belonging to the c-th dyneme class.

Conventional LDA assumes that the number of training samples N is adequately larger than the dimensionality of the input space, F, which is rarely the case. One way counteracting the problem is to first apply PCA to reduce the dimensionality of the data and

optimally preserve representational information in the least square sense, as proposed in [5]. PCA considers unlabelled data and seeks for the projection $\mathbf{W}_{pca} \in \Re^{F\times D}$ that maximizes the determinant of the total scatter matrix $\mathbf{S}_t$

$$\mathbf{S}_t = \sum_{\imath=1}^{N}(\mathbf{z}_\imath - \boldsymbol{\rho})(\mathbf{z}_\imath - \boldsymbol{\rho})^T , \quad (6)$$

$$\mathbf{W}_{pca} = \underset{\mathbf{W}}{\operatorname{argmax}} \mid \mathbf{W}^T\mathbf{S}_t\mathbf{W} \mid , \quad (7)$$

where $\boldsymbol{\rho}$ is the total mean of the weighted posture vectors, and T denotes matrix transposition. Therefore, the data are projected to yield a set of posture vectors in the projection space, $\{\mathbf{s}_1^{(1)}, \ldots, \mathbf{s}_{N_1}^{(1)}, \ldots, \mathbf{s}_{N_C}^{(C)}\}$, where $\mathbf{s} = \mathbf{W}_{pca}^T\mathbf{z}$.

The dimension of the projection space D is selected to retain most of the energy of the posture set in the input space. In the same time, assuming that $D < N-C$, LDA can be used to further project the posture vectors. The conventional LDA algorithm [2] seeks for the linear projection $\mathbf{W}_{lda} \in \Re^{D\times C-1}$ that maximizes the ratio of the between- and within-class scatter represented with the respective scatter matrices $\mathbf{S}_b$, $\mathbf{S}_w$ as outlined bellow

$$\mathbf{S}_b = \sum_{c=1}^{C} N_c(\boldsymbol{\mu}^{(c)} - \boldsymbol{\mu})(\boldsymbol{\mu}^{(c)} - \boldsymbol{\mu})^T , \quad (8)$$

$$\mathbf{S}_w = \sum_{c=1}^{C}\sum_{n=1}^{N_c}(\mathbf{s}_n^{(c)} - \boldsymbol{\mu}^{(c)})(\mathbf{s}_n^{(c)} - \boldsymbol{\mu}^{(c)})^T , \quad (9)$$

$$\mathbf{W}_{lda} = \underset{\mathbf{W}}{\operatorname{argmax}} \frac{\mid \mathbf{W}^T\mathbf{S}_b\mathbf{W} \mid}{\mid \mathbf{W}^T\mathbf{S}_w\mathbf{W} \mid} , \quad (10)$$

where $\boldsymbol{\mu}^{(c)}$ is the mean of the vectors in the c-th posture class and $\boldsymbol{\mu}$ is the total mean. The rank of $\mathbf{S}_w \in \Re^{D\times D}$ is at most $N-C$, and thus, is invertible if D has been adequately chosen. In this case, $\mathbf{W}_{lda}$ is formed from the generalized eigenvectors of $\mathbf{S}_w^{-1}\mathbf{S}_b$, and a vector $\mathbf{s} \in \Re^D$ is transformed using $\mathbf{p} = \mathbf{W}_{lda}^T\mathbf{s}$. Therefore, the final representation of a posture vector $\mathbf{x}$ in the feature space is given by

$$\mathbf{p} = \mathbf{W}_{opt}^T\mathbf{x} , \quad (11)$$

where $\mathbf{W}_{opt} = \mathbf{W}_{pca}\mathbf{W}_{lda}$.

3.3 Movement representation and classification

Let $\mathcal{U}$ be database of M movement sequences, where each sequence belongs to one of K different movement classes. A movement sequence of length L is projected to the dyneme classes using (11) to yield a sequence of posture vectors $\{\mathbf{p}_\ell\}_{\ell=1:L}$. Then, the sequence is partitioned to R parts of equal length, and the linear mean for each part is taken to represent the specific part

$$\underbrace{\mathbf{p}_1, \ldots, \mathbf{p}_{L_1}}_{\mathbf{q}_1}, \ldots, \underbrace{\mathbf{p}_{L_{R-1}+1}, \ldots, \mathbf{p}_{L_R}}_{\mathbf{q}_R}, \quad (12)$$

where $\mathbf{q}_r = \frac{1}{L_r}\sum_{\jmath=L_{r-1}+1}^{L_r}\mathbf{p}_\jmath$, and $L_0 = 0$, $L_R = L$. The resulted vectors are concatenated to represent the specific movement with the so-called *movement vector* $\mathbf{y} \in \Re^{RC}$

$$\mathbf{y} = [\mathbf{q_1}^T, \ldots, \mathbf{q_R}^T]^T . \quad (13)$$

Therefore the $\imath$-th movement sequence of the k-th movement class in the database is represented by the respective vector $\mathbf{y}_\imath^{(k)}$, yielding a set of movement vectors in the database $\{\mathbf{y}_1^{(1)}, \ldots, \mathbf{y}_{M_1}^{(1)}, \ldots, \mathbf{y}_{M_K}^{(K)}\}$, where, M_k is the number of movements in the k-th class. Note that LDA can again applied to project the movement vectors in $\Re^{K-1}$ and further enhance class discrimination and compactness of representation.

Assuming that the movement classes are derived from unimodal gaussian distributions with the same covariance matrix $\boldsymbol{\Sigma}$ but different means $\boldsymbol{\eta}^{(k)}$, $k = 1, \ldots, K$, we can use the maximum likelihood technique to estimate them

$$\boldsymbol{\eta}^{(k)} = \frac{1}{M_k}\sum_{\imath=1}^{M_k}\mathbf{y}_\imath^{(k)}, \quad k = 1, \ldots, K , \quad (14)$$

$$\boldsymbol{\Sigma} = \frac{1}{M}\sum_{\imath=1}^{M}(\mathbf{y}_\imath - \boldsymbol{\eta})(\mathbf{y}_\imath - \boldsymbol{\eta})^T , \quad (15)$$

where $\boldsymbol{\eta}$ is the total mean ($\boldsymbol{\eta} = \frac{1}{M}\sum_{\imath=1}^{M}\mathbf{y}_\imath$).

A novel movement sequence is projected with (11) and represented with a movement vector $\mathbf{e}^{(b)} \in \Re^{RC}$ using (12),(13). Assuming equiprobable priors, the class label b of the novel movement is given by:

$$b = \underset{k\in[1,\ldots,K]}{\operatorname{argmin}} (g_k(\mathbf{e})) , \quad (16)$$

where g_k, $k = 1, \ldots, K$, are the discriminant functions

$$g_k(\mathbf{e}) = (\mathbf{e} - \boldsymbol{\eta}^{(k)})^T\boldsymbol{\Sigma}^{-1}(\mathbf{e} - \boldsymbol{\eta}^{(k)}) \quad (17)$$

Alternatively, the maximum cosine distance can be used to classify the novel movement

$$g_k(\mathbf{e}) = \frac{-\mathbf{e}^T\boldsymbol{\eta}^{(k)}}{\parallel \mathbf{e} \parallel \parallel \boldsymbol{\eta}^{(k)} \parallel} . \quad (18)$$

3.4 Multiview movement recognition

The proposed method can be extended for view independent human movement recognition, exploiting a multicamera infrastructure during the training stage, similar to [11]. That is, assuming P cameras, a movement model, $\boldsymbol{\eta}^{(k,p)}$, is build for each movement k and each viewing angle p, i.e., $\boldsymbol{\eta}^{(k,p)}, k = 1, \ldots, K$, $p = 1, \ldots, P$, and the respective discriminant function $g_{k,p}(\cdot)$ is used, as described in the previous section. Then a novel movement vector $\mathbf{e}^{(b)}$, captured from an arbitrary view, is classified according to the label of the closest movement model.

4. EXPERIMENTAL RESULTS

The classification database reported in [7] is used to assess the performance of the method. This database contains nine persons performing ten movements, namely,

"walking" (wk), "running" (rn), "skipping" (sk), "galloping sideways" (sd), "jumping jack" (jk), "jumping" (jp), "jumping in place" (pj), "bending" (bd), "wave with one hand" (wo) and "wave with two hands" (wt). The database contains in total 93 videos (3 persons perform the same movement 2 times). The proposed algorithm assumes that each video contains only a single instance of a movement. In contrary, some videos show a person executing several cycles of a periodic movement. We brake such videos to their constituting single movements, and thus, we produce a database of 230 movement instances. The movement videos comprise variable inter- and intra-class length, for instance, the smallest movement of "run" consists of 9 frames, while the largest video of "bend", consists of 66 frames respectively.

Next, we transform movement sequences to show persons moving in the same direction, either left or right. This is done by first deciding the direction, and then mirroring the frames of the movement videos that show a person moving to an opposite direction from the one decided. Movement direction detection is done automatically by observing the displacement of the mask on the x frame axis over a few frames in time.

In our computations, the associated binary masks of the classification database are employed. From each binary mask, the rectangular region containing the silhouette is extracted, to form a silhouette image. All silhouette images along a movement video are centered according to the center of mass of the silhouette at each image. Then, each silhouette image is transformed to the same size, here 64×48, with bicubic interpolation (as in [6]), to represent a movement video with a sequence of posture images. (For instance, two foreground posture images of "skip" and two of "run" are depicted in Fig. 1. Here we use the respective binary masks.) The resulted images are then scanned column-wise to form 3072-dimensional posture vectors.

The leave-one-out-cross-validation (LOOCV) procedure is used to assess the performance of the algorithm. At each validation cycle all the movement sequences referring to a specific person performing a specific movement are extracted to form the test set. The remaining movement sequences are used as a training sequence to compute the movement prototypes as described in section 3.3. The number of corrected classified movement sequences at each cycle are summed to compute the classification rate. Extensive experiments have been

Figure 2: *Three dynemes identified with the proposed method. Dyneme 14 resembles a posture of "bend", while dyneme 33 and 17 a posture of "walk" and "wave with two hands" respectively. Dyneme 33 could as well represent a posture of "run".*

performed to identify the number of clusters C, fuzzification parameter m, and movement partition number R. For $C = 43$, $m = 1.14$ and $R = 2$ a recognition rate of 90.4% was achieved, i.e., only 22 out of 230 movement sequences were misclassified. Some of the identified dynemes clearly characterize the movements, while other are confused between two or more movements. For instance, in Fig. 2, dyneme 13 clearly characterize the movement of bend, while dyneme 33 is confused between the posture of walk and run. The confusion matrix for these settings is shown in Table 1. In this table we see

	bd	jk	jp	pj	rn	sd	sp	wk	wo	wt
bd	9									
jk		17								1
jp			20	1	1	3	4			
pj				25						
rn					27			1		
sd				1		18	1		2	
sp			2				27			
wk								42		
wo					1				13	
wt									4	10

Table 1: *Confusion between movements. A row represents the actual movement and the column the name of the movement recognized by the algorithm during the LOOCV procedure.*

that "jump" is the movement that confused mostly. On the other hand, "bend", "jumping in place" and "walk" are well recognized by the proposed algorithm.

The recognition rate achieved here is better than the rate attained with the real time Dominant sets-based method reported in [8]. In this work, the number of clusters are identified using Dominant Sets. Although the clusters identified there may well represent the intrinsic structure of the input space, it is not guaranteed that they provide the dyneme that optimally discriminate different movements, as we pursue in our method.

Apart from Dominant Sets, FCM as well as other clustering algorithms were also applied in [8], to represent each movement class with a cluster centroid, and recognize a novel movement according to nearest centroid distance. The attained classification results of this approach were not satisfying. In our method, FCM is explicitly used to identify the dyneme classes, which uniquely characterize different movements, and express movements upon dynemes. Thus, classification accuracy is considerable improved.

The Carnegie Mellon University motion capture database [12] contains motion capture data of several persons performing several different movements. We used this database to synthesize artificial multiview binary mask sequences and test the applicability of the algorithm within a multiview scenario. Initial results on this direction are promising, showing that the proposed method can be used for multiview human movement recognition.

5. CONCLUSION

A novel human movement representation and recognition method has been proposed. A movement of any length is compactly expressed in terms of its compris-

ing dynemes, as a single vector in a low dimensional space. This representation allows simple cosine- or Mahalanobis-comparison of different movements, avoiding expensive comparison metrics, and thus, offering higher speed and storage efficiency from prevailed methods in the field, e.g., [6], [7], [10]. Moreover, the recognition rate achieved here (90.4%) on a publicly available database, outperforms rates reported in other real-time methods in the same database, e.g., [8].

A possible extension of the method for multiview human movement recognition has also be presented. Initial experimental results on this direction were promising and will be reported in a future publication.

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 211471 (i3DPost).

REFERENCES

[1] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Address: Plenum, New York, 1981.

[2] R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification, 2nd ed.*. Address: John Wiley and Sons, 2001.

[3] T. B. Moeslund, A. Hilton, V. Kruger, "A survey of advances in vision-based human motion capture and analysis," *Comput. Vis. Image Understand.*, vol. 104, no. 2, pp. 90–126, Nov. 2006.

[4] R.D. Green, L. Guan, "Quantifying and recognizing human movement patterns from monocular video Images-part I: A new framework for modeling human motion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 2, pp. 179–190, Feb. 2004.

[5] P.N. Belhumeur, J.P. Hespanha, D.J Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[6] L. Wang, D. Suter, "Learning and matching of dynamic shape manifolds for human action recognition," *IEEE Trans. on Image Proc.*, vol. 16, no. 6, pp. 1646–1661, Jun. 2007.

[7] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri, "Actions as Space-Time Shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.29, no. 12, pp. 2247–2253, Dec. 2007.

[8] Q. Wei, W. Hu, X. Zhang, G. Luo, "Dominant Sets-Based Action Recognition using Image Sequence Matching " in *Proc. IEEE Int. Conf. Im. Proc. 2007*, San Antonio, TX, USA, September 2007, vol. 6, pp. 1522–4880.

[9] A. F. Bobick, J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

[10] J. Ben-Arie, Z. Wang, P. Pandit, S. Rajaram, "Human activity recognition using multidimensional indexing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 8, pp. 1091–1104, Aug. 2002.

[11] M. Ahmad, S.W. Lee, "HMM-based Human Action Recognition Using Multiview Image Sequences," *Proc. 18th Int. Conf. on Pattern Recognition*, vol. 1, pp. 263–266 , Aug. 2006.

[12] "Carnegie Mellon University Motion Capture Database," `http://mocap.cs.cmu.edu/`.