A COMPREHENSIVE ANALYSIS OF HUMAN MOTION CAPTURE DATA FOR ACTION RECOGNITION

Valsamis Ntouskos, Panagiotis Papadakis and Fiora Pirri
ALCOR, Vision, Perception and Cognitive Robotics Laboratory, Department of Computer and System Sciences,
University of Rome "La Sapienza", Rome, Italy

Keywords: Motion Analysis, Motion Capture, Action Recognition.

Abstract: In this paper, we present an analysis of human motion that can assist the recognition of human actions irrespective of the selection of particular features. We begin with an analysis on the entire set of preclassified motions in order to derive the generic characteristics of articulated human motion and complement the analysis by a more detailed inter-class analysis. The statistical analysis concerns features that describe the significance-contribution of the human joints in performing an action. Furthermore, we adopt a hierarchical analysis on the human body itself in the study of different actions, by grouping joints that share common characteristics. We present our experiments on standard databases for human motion capture data as well as a new commercial dataset with additional classes of human motion and highlight certain interesting results.

1 INTRODUCTION

Research interest has been largely stimulated by the analysis of human motion as it constitutes a key component in a plurality of disciplines. Common applications range from human action recognition, human-machine interaction, skill learning and smart surveillance to applications within the entertainment industry such as character animation, computer games and film production.

As manifested by earlier research (Mihai, 1999), (Kovar and Gleicher, 2004), (Thomas et al., 2006), (Poppe, 2010) the study of human motion is not a new research field, however, most of the focus has so far been directed towards 2D image-based human motion representations. In contrast, with the evolution of motion capture hardware-software and recently with the advent of affordable depth acquisition devices such as the Kinect (Microsoft, 2010), part of research is shifting towards 3D (spatial) representations of human motion extracted from a hierarchical representation of the human pose, i.e. a skeletal structure.

Motion capture data (MOCAP) contain information of the human pose as recorded during the execution of an action that is organized into a collection of 3D points-joints together with the spatial position or rotation of the corresponding coordinate frames. These points comprise a skeletal representation of the human pose wherein the motion of the points has been acquired either directly from body sensors (optical or magnetic) or tracked along the action sequence.

Previous studies on the characteristics of articulated human motion such as (Fod et al., 2002), (Pullen and Bregler, 2002), (Jenkins and Mataric, 2003), (Barbič et al., 2004), (Okan, 2006) have been conducted within various contexts, namely, 3D animation compression, motion synthesis and motion indexing-classification or segmentation. However, experiments are usually performed on relatively small collections of human motions with relatively limited variability in the classes of motion.

In this work, we provide a comprehensive study of human motion, through a statistical analysis of a diverse set of human action categories using MOCAP databases. Our experiments are performed on standard databases (CMU, 2003), (Muller et al., 2007) and further extended on a commercial database (mocapdata.com, 2011), altogether highlighting important characteristics of articulated human motion related to the correlation of human motion and hierarchy of the human kinematic chain in terms of body part contribution. In particular, we perform a statistical analysis on various abstraction levels of the joint representation of the human skeleton across different datasets and investigate the variance within the inherent dimensionality of simple and complex actions. The motivation of our study is to provide a deeper insight into the characteristics of articulated human mo-

Ntouskos V., Papadakis P. and Pirri F..
A COMPREHENSIVE ANALYSIS OF HUMAN MOTION CAPTURE DATA FOR ACTION RECOGNITION.
DOI: 10.5220/0003868806470652
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2012), pages 647-652
ISBN: 978-989-8565-03-7

tion in order to assist the process of machine recognition of human actions.

The remainder of the paper is organized as follows: in Section 2 we unfold the necessary formulations with respect to the experimentation set-up, in Section 3 we present the results of the human motion analysis within various datasets and across different classes and finally, in Section 4, we summarize the conclusions.

2 HUMAN MOTION REPRESENTATION

Commonly used representations for human motion capture include the ASF/ACM, C3D and the BVH file format (adopted in this work) that describes in plain text the structure of the actor's skeleton along with the data acquired during the motion capture. In order to avoid inconsistencies in the order of rotation angles, for each joint within the BVH structure, we obtain a structure built on top of the corresponding unit norm quaternions. In detail, a quaternion $q \in \mathbb{H}$ is specified as follows:

$$q = (a + ib + jc + kd) \tag{1}$$

where $(a,b,c,d) \in \mathbb{R}$ and i,j and k form the basis of $\mathbb{H}$. The quaternion norm is given by $||q|| = \sqrt{qq^\star}$, where $q^\star = (a - ib - jc - kd)$ is the quaternion conjugate. We use unit-norm quaternions q_w to indicate the rotation of the w-th joint by an angle θ around a unit vector $\hat{\mathbf{u}}$:

$$q_w = (\cos\left(\frac{1}{2}\theta\right), \hat{\mathbf{u}}\sin\left(\frac{1}{2}\theta\right)) \tag{2}$$

An action A is described by n frames F_j, $j = 1,\dots,n$, wherein each frame F_j of the motion capture sequence captures the pose of the human skeleton (as shown in Figure 1) that consists of m joints .

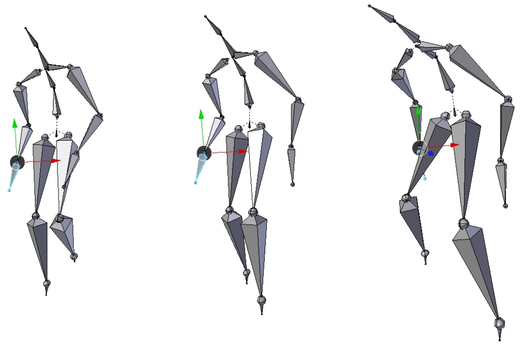

Figure 1: The skeletal representation of human joints for a subset of frames of a MOCAP sequence.

We note here that for our experimental analysis, we have considered only actions for which $n > 4m$, that is, only action sequences with sufficient time duration. Our goal is to study inter as well as intra-class variations in the dimensionality of human actions, in order to establish the conditions for classification.

The first step is to obtain a covariance matrix of the pose for each frame F_j based on the quaternion structure. To this end we follow both (Cheong Took and Mandic, 2011) and (Ginzberg and Walden, 2011) to establish a correspondence between the domain of a quadrivariate random variable in $\mathbb{R}^4$ and the quaternion domain $\mathbb{H}$.

Let us consider the real components of the quaternion q as $q_a = [a,b,c,d]$. Then, for a given pose in a frame F_j, for example of m joints, we can set the real quaternion values of the joints as real-valued m-dimensional jointly Gaussian random variables as follows.

Let the values $\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d}$ be m-dimensional columns vectors, then for each frame F_j we obtain a $4m$ dimensional Gaussian vector:

$$\mathbf{r} = \left[\mathbf{a}^\top, \, \mathbf{b}^\top, \, \mathbf{c}^\top, \, \mathbf{d}^\top\right] \tag{3}$$

The real covariance $\mathbf{E}(\mathbf{r}\mathbf{r}^\top) = \Sigma_r$ of this vector is defined as usual:

$$\Sigma_r = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} & \Sigma_{ac} & \Sigma_{ad} \\ \Sigma_{ab} & \Sigma_{bb} & \Sigma_{bc} & \Sigma_{bd} \\ \Sigma_{ac} & \Sigma_{bc} & \Sigma_{cc} & \Sigma_{cd} \\ \Sigma_{ad} & \Sigma_{bd} & \Sigma_{cd} & \Sigma_{dd} \end{pmatrix} \tag{4}$$

Σ_r is a $4m \times 4m$ matrix of the covariance of all the real valued quadrivariate vectors in $\mathbb{R}^4$ associated with the quaternions specified by the m joints in the pose of the action at F_j. Looking at Σ_r as a block matrix then Σ_{xy} is a $m \times m$ matrix. In particular, if we subtract the mean value μ_r from $\mathbf{r}$ then, from the eigen decomposition of $\Sigma_r = \mathbf{U}\Lambda\mathbf{U}^\top$, we obtain the $4m$ directions of the hyper-ellipsoid specified by Σ_r. In particular, $\mathbf{U}\mathbf{U}^T = \mathbf{I}$, Λ is the matrix of positive eigenvalues, returning the length of the axes of the hyper ellipsoid, namely, the length l_i of the i-th hyper ellipsoid axis is $l_i = 2\sqrt{\lambda_i}$, λ_i the i-th eigenvalue of Λ, and its direction is specified by $\mathbf{u_i}$ the i-th eigenvector of $\mathbf{U}$.

Since the hyper ellipsoid is centered, according to the transformation, the n-variate Gaussian distribution associated with each *centered* pose of action A can be specified as follows, where $\mathbf{x}$ is a real valued vector $4m \times 1$ of a pose:

$$\mathcal{N}(\mathbf{x}|\mathbf{0}_{4m\times1},\Sigma_r) = \frac{1}{2\pi^{n/2}|\Sigma_r|^{1/2}} \exp\{-\frac{1}{2}\mathbf{x}^\top\Sigma_r^{-1}\mathbf{x}\} \tag{5}$$

Therefore several distance transforms can be used to verify both intra and inter classes distances, like the

Mahalanobis distance or the Kullback-Leibler divergence for the Gaussian pdf related to different poses or different actions.

According to (Cheong Took and Mandic, 2011) and (Ginzberg and Walden, 2011) the second-order information carried by the random vector quaternion valued $\mathbf{q}$ are not fully specified by the above covariance matrix, hence the following transformation is introduced:

$$A = \begin{pmatrix} \mathbf{I} & i\mathbf{I} & j\mathbf{I} & k\mathbf{I} \\ \mathbf{I} & i\mathbf{I} & -j\mathbf{I} & -k\mathbf{I} \\ \mathbf{I} & -i\mathbf{I} & j\mathbf{I} & -k\mathbf{I} \\ \mathbf{I} & -i\mathbf{I} & -j\mathbf{I} & k\mathbf{I} \end{pmatrix} \tag{6}$$

Here $\mathbf{I}$ is a $m \times m$ identity matrix. However to obtain the transformation the quaternion vector needs to be augmented with its involutions, namely self-inverse mapping $-iqi$,$-jqj$, and $-kqk$ (see (Cheong Took and Mandic, 2011)), thus $\mathbf{q} = A\mathbf{r}$, where $\mathbf{q} = [\mathbf{q}, \mathbf{q}^i, \mathbf{q}^j, \mathbf{q}^k]$ and $q^i = -iqi$, $q^j = -jqj$ and $q^k = -kqk$. Then the covariance of a m dimensional Gaussian random variable $\mathbf{q}$ is:

$$\begin{aligned} &\Sigma_{\mathbf{q}} = E(\mathbf{q}\mathbf{q}^{\mathbf{H}}) = A\Sigma_r A^H \\ &\text{conversely} \\ &\Sigma_r = \frac{1}{16} A^H \Sigma_{\mathbf{q}} A \end{aligned} \tag{7}$$

Here $\mathbf{q}^H$ is the conjugate transpose of $\mathbf{q}$. We note that the covariance $\Sigma_{\mathbf{q}}$ of the random variate $\mathbf{q}$ comprise the above mentioned involutions. It turns out that a quaternion is not correlated with its vector involutions, hence the covariance $\Sigma_{\mathbf{q}}$ can be simplified yielding $\Sigma_q = diag\{\Sigma_{qq}, \Sigma_{q^i q^i}, \Sigma_{q^j q^j}, \Sigma_{q^k q^k}\}$, $\Sigma_{qq} = E\{\mathbf{q}\mathbf{q}^H\}, \Sigma_{q^i} = E\{\mathbf{q}\mathbf{q}^{iH}\}$, $\Sigma_{q^j} = E\{\mathbf{q}\mathbf{q}^{jH}\}$ and $\Sigma_{q^k} = E\{\mathbf{q}\mathbf{q}^{kH}\}$. Therefore, finally, the Mahalanobis distance for a multivariate quaternion-valued random vector $\mathbf{q}$, augmented with its involutions, is:

$$\mathbf{q}^H \Sigma_{\mathbf{q}}^{-1} \mathbf{q} \tag{8}$$

For each pose of any action A we have obtained the covariance parameter of the $4m$-variate Gaussian random variate associated with it, both in terms of the real values of the pose and of its augmented quaternion values.

Now, it is easy to see that the above representation of a pose is independent of the length of an action, and it depends only on the number of joints, namely on the pose at each frame F_j. Clearly this turns out to be a problem only for comparison between different data sets, as in the same data sets all poses have the same number of joints.

Therefore, in order to suitably compare actions A_k, $k = 1, ..,$ across different datasets it is necessary to align two poses with respect to the number of joints. In order to obtain the number of joints whose quaternions have to be combined we proceed as follows.

First we compute the distribution of joints across two poses, say X and Y and then we compute the average quaternion about the exceeding joints. We recall that the average quaternion can be estimated by solving the following optimization problem (Markley et al., 2007):

$$\begin{aligned} &\bar{\mathbf{q}} = \underset{\mathbf{q} \in \mathbb{S}^3}{\text{argmax}}\ \mathbf{q}^\top \mathbf{M} \mathbf{q} \\ &\text{with: } \mathbf{M} = \sum_1^n w_i \mathbf{q}_i \mathbf{q}_i^\top \text{ and} \\ &\mathbb{S}^3 \text{ denoting the unit 3-sphere} \end{aligned} \tag{9}$$

A solution to this problem can be computed by taking the eigenvector of $\mathbf{M}$ which corresponds to its largest eigenvalue. All weights w_i were taken equal to 1 for the successive analysis. In this process joints corresponding to fingers and toes were not considered mainly due to the high noise level which degraded the quality of these joint measurements in all datasets. These average quaternions have been successively treated analogously to what has been shown previously by computing the covariance matrix and performing its eigen decomposition.

Let J be the number of joints of a pose X of action A_i and let W be the number of joints of a pose Y of action B. Assume that $J < W$ (the same reasoning can be applied in the inverse case), then let $\delta j = W/J$, and let:

$$C = \lceil (1, \ldots, J)\delta j \rceil \tag{10}$$

Here $\lceil \cdot \rceil$ is the ceiling operator, and C is the cumulative number of joints that should be combined so as to map W joints onto J joints. Further, to obtain the inverse cumulative operation C^{-1} we define

$$C^\star = [C_1,\ C_2 - C_1, \ldots, C_{W-1} - C_{W-2}]^\top \tag{11}$$

Hence the length of both C and $C^\star$ is J. It is easy to see that:

$$\sum C^\star = W \tag{12}$$

Now, for each joint j, the graph of its connecting joints is defined, then C tells which joint j is interested in the average, and $C^\star$ tells how many joints of Y needs to be averaged at j, so as two obtain two poses with the same number of joints.

For example if $J = 7$ and $W = 19$ then $C = (3, 6, 9, 11, 14, 17, 19)^\top$ and $C^\star = (3, 3, 3, 2, 3, 3, 2)^\top$, hence at the joint $j = C(4) = 11$, we have that $C^\star(4) = 2$, hence two joints of the graph of $j = 11$ will be averaged as indicated in (9).

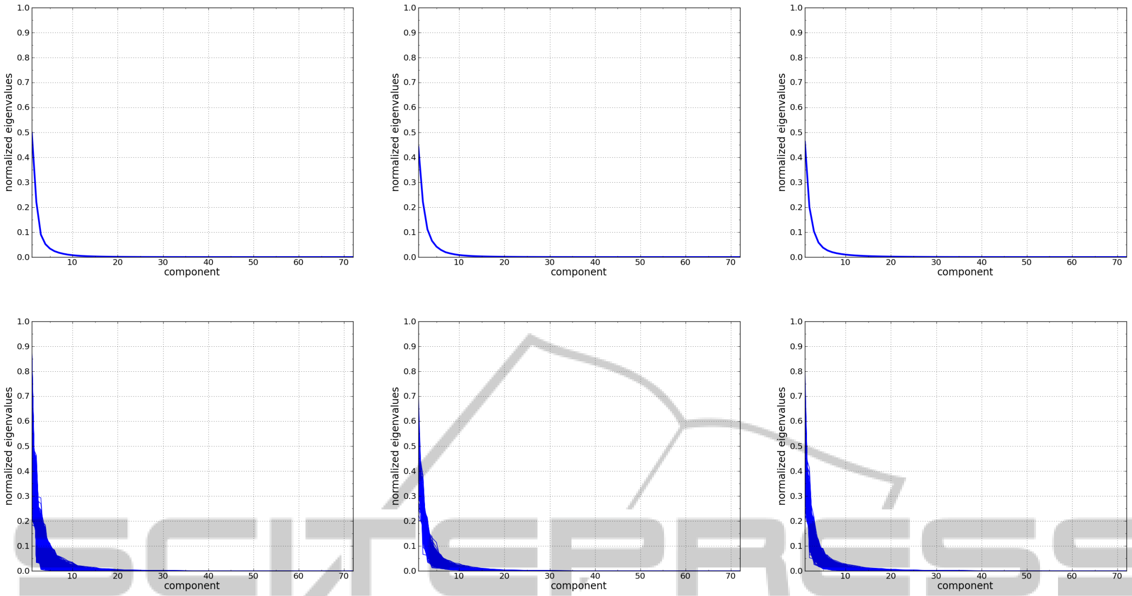

Figure 2: Top row: Eigenvalues of the CMU, HDM05 and MEJ datasets respectively. Bottom row: Eigenvalues of all the sequences of the CMU, HDM05 and MEJ datasets respectively.

3 EXPERIMENTS

In this section we present our experiments regarding the analysis of the human motion across different MOCAP datasets and highlight the significance of the results in the context of action recognition. The analysis was performed on the following, publicly available, MOCAP datasets:

- The *Carnegie Mellon University* (**CMU**) dataset (CMU, 2003).
- The *Hochschule der Medien* (**HDM05**) dataset (Muller et al., 2007).
- The *Mocapdata Eyes Japan* (**MEJ**) dataset (mocapdata.com, 2011).

whose characteristics are given in Table 1:

Table 1: Characteristics of MOCAP datasets used in the experiments.

Dataset	actors	classes	motions	fps	joints
CMU	144	23	2605	120	31
HDM05	5	28	293	120	31
MEJ	3	18	675	30	19

The classes of human motion within the datasets capture an extensive range of human actions, from common and simple actions such as walking and jumping to more complex such as sporting, dancing, martial arts, manipulation actions and gestures.

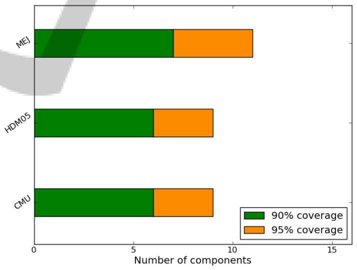

Figure 3: Covariance coverage w.r.t. the number of dimensions considered.

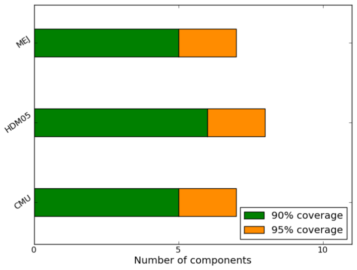

Figure 4: Limbs level: Covariance coverage w.r.t. the number of dimensions considered.

The motion sequences differ in the number of skeletal joints considered as well as the frame rate, however the results are not biased by these factors.

In Figure 2 the covariances of all the individual motion sequences as well as the resulting mean eigen-

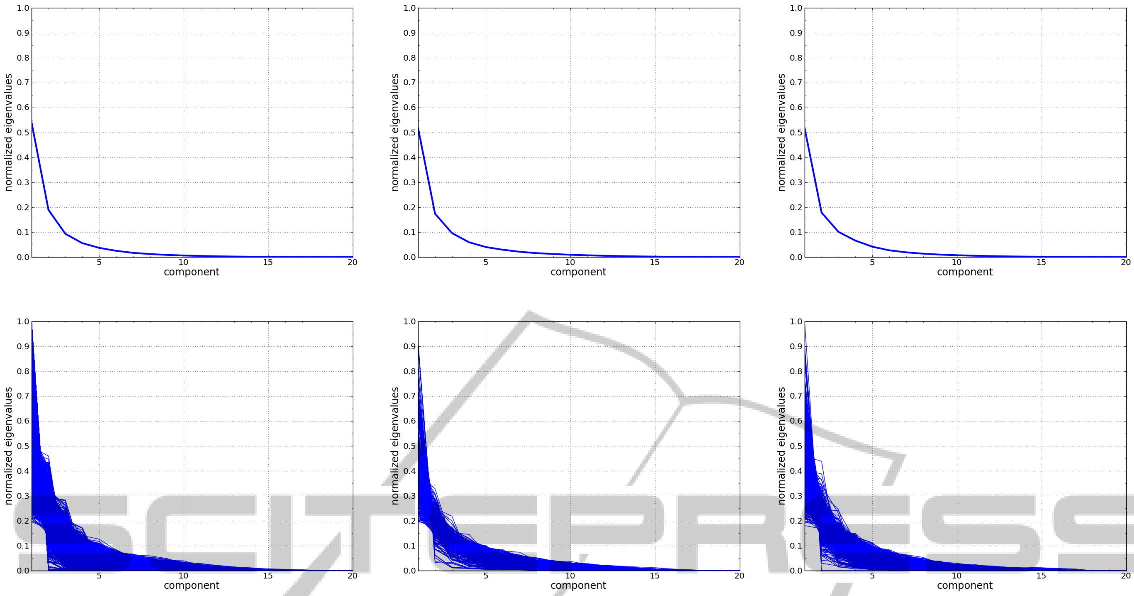

Figure 5: **Limbs level**: Top row: Average eigenvalues of the CMU, HDM05 and MEJ datasets respectively. Bottom row: Eigenvalues of all the individual sequences of the CMU, HDM05 and MEJ datasets respectively.

values of the real valued poses $\mathbf{q}_a$ (see Section 2) is presented for the three aforementioned datasets.

It is evident that the variances of the motion sequences follow a specific pattern irrespectively of the particular action being performed. In particular, we observe a high degree of correlation of human motion that is quantitatively demonstrated in Figure 3, that presents the number of dimensions necessary to cover 90% and 95% of the pose variances for each dataset. Figure 4 shows the number of dimensions necessary to cover 90% and 95% of the variance, by grouping joints that belong to the same skeletal limb.

Figure 5 shows the variance of the motion sequences for each dataset by performing the analysis on the limbs level. One can note that also in this case a small portion of the overall dimensions carries most of the information with the eigenvalues decreasing exponentially like before.

Besides considering motion sequences from different datasets we have also analysed the variances and the real valued eigenvalues of sequences of different actions taken from the same dataset. For each of the datasets, Figure 6 shows the two actions for which the respective eigenvalues decay with the lowest rate, in other words, the most complex actions, as well as the two actions for which the eigenvalues decay with the highest rate, i.e. the simplest actions. Figure 6 also presents the average eigenvalues obtained from all the sequences of the respective dataset for comparison. Semi-logarithmic scaling has been used in order to highlight the differences in the eigenvalues distribution. Finally, in Figure 7, we show the number of dimensions necessary to cover 90% and 95% of the variances for the respective action classes within each dataset. These results indicate that, as intuitively expected, the correlation in the dimensions describing a human motion is lower for more complex actions and higher for simple actions like gestures or manipulation of objects. However, even for the most complex actions, the necessary dimensionality remains relatively low.

4 CONCLUSIONS

In this work we have considered several motion sequences taken by various available datasets. As has been experimentally shown, an analysis of the correlation of the motion sequences shows that the majority of the information regarding the human motion resides in a lower dimensional space. Moreover, the trend by which the amount of information carried by the components is decreasing in the fine resolution level (considering all the joints) is similar to the coarse resolution level, i.e. at the limb level. Finally, we also highlighted that individual actions which are considered as more semantically ”complex” give a spectrum that decays with a lower rate compared to the one resulting from ”simpler” actions. These considerations further support the argument that human

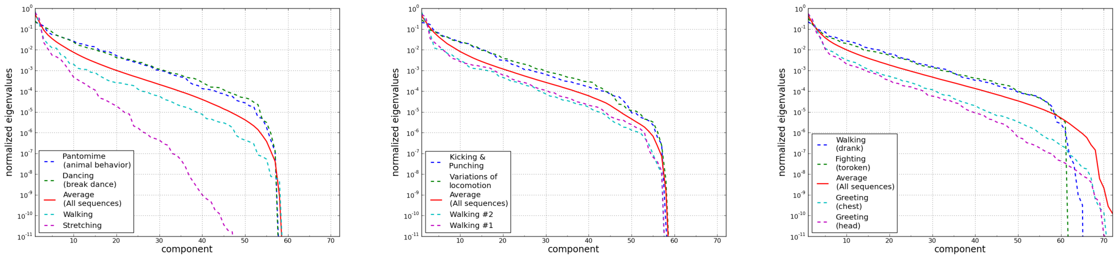

Figure 6: Eigenvalues of different actions for the CMU, HDM05 and MEJ datasets respectively (semilogarithmic scale).

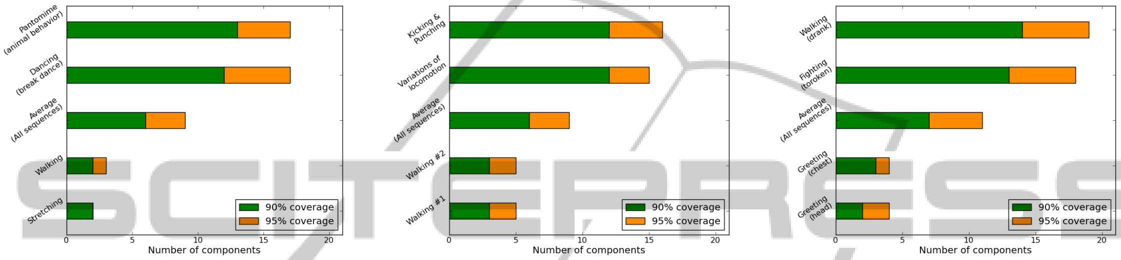

Figure 7: Covariance coverage of different actions for the CMU, HDM05 and MEJ datasets respectively.

motion can be classified using a representation which considers a relatively low number of dimensions.

ACKNOWLEDGEMENTS

We would like to thank mocapdata.com, Eyes, JAPAN Co. Ltd. for providing the full bundle of the MOCAP sequences. This paper describes research done under the EU-FP7 ICT 247870 NIFTI project.

REFERENCES

Barbič, J., Safonova, A., Pan, J.-Y., Faloutsos, C., Hodgins, J., and Pollard, N. (2004). Segmenting motion capture data into distinct behaviors. In *Proc. Graphics Interface 2004*, pages 185–194.

Cheong Took, C. and Mandic, D. P. (2011). Augmented second-order statistics of quaternion random signals. *Signal Process.*, 91:214–224.

CMU (2003). Carnegie-mellon mocap database. http://mocap.cs.cmu.edu/.

Fod, A., Matari, M., and Jenkins, C. (2002). Automated derivation of primitives for movement classification. *Autonomous Robots*, 12:39–54.

Ginzberg, P. and Walden, A. (2011). Testing for quaternion propriety. *Signal Processing, IEEE Transactions on*, 59(7):3025 –3034.

Jenkins, C. and Mataric, M. (2003). Automated derivation of behavior vocabularies for autonomous humanoid motion. In *AMAS*, pages 225–232. ACM.

Kovar, L. and Gleicher, M. (2004). Automated extraction and parameterization of motions in large data sets. In *ACM SIGGRAPH 2004*, pages 559–568.

Markley, F. L., Yang, C., Crassidis, J. L., and Oshman, Y. (2007). Averaging quaternions. *Guidance, Control, and Dynamics*, 30(4):1193 – 1196.

Microsoft, C. (2010). Kinect. http://www.xbox.com/en-US/kinect.

Mihai, G. D. (1999). The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82 – 98.

mocapdata.com (2011). Eyes, japan co. ltd. http://www.mocapdata.com/.

Muller, M., Roder, T., Clausen, M., Eberhardt, B., Krüger, B., and Weber, A. (2007). Documentation mocap database hdm05. Technical Report CG-2007-2, Universität Bonn.

Okan, A. (2006). Compression of motion capture databases. *ACM Transactions on Graphics*, 25:890–897.

Poppe, R. (2010). A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976 – 990.

Pullen, K. and Bregler, C. (2002). Motion capture assisted animation: texturing and synthesis. In *Proc. of the 29th annual conf. on Computer graphics and interactive techniques*, pages 501–508. ACM.

Thomas, M., Adrian, H., and Volker, K. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2-3):90 – 126.