Optimizing orthonormal basis bilinear spatiotemporal representation for motion data

Przemysław Skurowski[1], Jolanta Socała[2], and Konrad Wojciechowski[3]

[1] Institute of Informatics, Silesian University of Technology, Gliwice, Poland, e-mail: przemyslaw.skurowski@polsl.pl, WWW: http://inf.polsl.pl
[2] Department of Mathematics, University of Bielsko-Biała, Bielsko-Biała, Poland, e-mail: socala.jolanta@gmail.com, WWW: http://www.km.ath.bielsko.pl
[3] Polish-Japanese Academy of Information Technology, Bytom, Poland, e-mail: konrad.wojciechowski@polsl.pl, WWW: http://www.bytom.pjwstk.edu.pl/

Abstract. The paper describes an attempt to estimate the optimal division of a number of base vectors between space (shape) and time (trajectory) for bilinear spatio-temporal representation of motion capture data. The spatiotemporal model is a matrix consisting of $K_s \cdot K_t$ amount of coefficients. In the paper we discuss using of orthonormal spatial and temporal basis: PCA-PCA, DCT-DCT, PCA-DCT and DCT-PCA to represent real MoCap data.

Keywords: spatiotemporal representation, motion capture, motion data, bilinear model

1 Introduction

Optical motion capture (Mocap) data [7, 8] comprises stored 3D trajectories of certain points over a time. The bilinear spatiotemporal representation [3] for motion data is a novel concept which offered an opportunity to store the motion information as spatiotemporal matrix being very efficient and compact form. The bilinear representation method can be considered as a combination of both spatial (shape) and temporal (trajectory) bases. The preferred bases are orthonormal ones, of which these obtained with principal component analysis (PCA) and from the discrete cosine transform (DCT) are most appreciated for a shape and trajectory respectively.

The model is very general as there are no limitations, so both rigid and soft objects can be represented. As it was demonstrated by the authors of the method, the representation appeared to be effective for denoising and labeling of non-rigid (facial) motion capture data. The model was also used to identify roles of football players [6, 12] by analysis of the team shape and motion of players.

In the paper, we address the problem of the optimal number of spatial and temporal dimensions which still remain open. We study all the four combinations of PCA and DCT basis for the reconstruction root mean square error (RMSE) for the bases with significantly reduced number of dimensions. We are looking for

the optimal sharing of a number of basis vectors between shape and trajectory for approximately constant (fixed) overall number of coefficients. The experiments were conducted for a number of MoCap sequences acquired mostly in the Human Motion Lab of Polish Japanese Academy of Information Technology.

2 The Method

2.1 Bilinear Spatiotemporal Basis

Assume we have the time-varying structure of a set of P points sampled at F time instances. It can be represented as a sequence of 3D points:

$$\mathbf{S}_{F\times 3P} = \begin{bmatrix} \mathbf{X}_1^1 & \dots & \mathbf{X}_p^1 \\ \vdots & & \vdots \\ \mathbf{X}_1^F & \dots & \mathbf{X}_p^F \end{bmatrix}, \tag{1}$$

where: $\mathbf{X}_j^i = [X_j^i, Y_j^i, Z_j^i]$ denotes the $3D$ coordinates of the j-th point at the i-th time instance. Obviously, the time-varying structure matrix $\mathbf{S}$ contains $3FP$ parameters. We indicate row-index as superscript and column-index as subscript.

We can represent the $3D$ shape at each time instance as a linear combination of a small number K_s ($K_s \ll 3P$) of shape basis vectors $\mathbf{b}_j$ weighted by coefficients ω_j^i [5, 4],

$$\mathbf{s}^i = \sum_j \omega_j^i \mathbf{b}_j^T. \tag{2}$$

Every shape basis vector represents a $3D$ structure of length $3P$. The structure matrix $\mathbf{S}$ can be represented as:

$$\mathbf{S} = \mathbf{\Omega}\mathbf{B}^T, \tag{3}$$

where: $\mathbf{B}$ is a $3P \times K_s$ matrix containing K_s shape basis vectors as its rows and $\mathbf{\Omega}$ is a $F \times K_s$ matrix containing the corresponding shape coefficients ω_j^i.

We have also another representation. We can represent every trajectory as a linear combination of a small number K_t ($K_t \ll F$) of trajectory basis vectors θ_i weighted by coefficients a_i^j [11, 2],

$$\mathbf{s}_j = \sum_i a_i^j \theta_i. \tag{4}$$

Every trajectory basis vector represents a structure of length F. The structure matrix $\mathbf{S}$ can be represented as:

$$\mathbf{S} = \mathbf{\Theta}\mathbf{A}^T, \tag{5}$$

where: $\mathbf{\Theta}$ is a $F \times K_t$ matrix containing K_t trajectory basis vectors and $\mathbf{A}$ is a $3P \times K_t$ matrix containing the corresponding trajectory coefficients a_i^j.

We will use the third method - the Bilinear Spatiotemporal Basis. We assume: the vectors $\mathbf{b_j}$ are orthonormal and the vectors $\theta_\mathbf{i}$ are orthonormal too. The following theorem [3] give us a bilinear representation of $\mathbf{S}$.

Theorem 1. *Let us assume that we have* $\mathbf{S} = \mathbf{\Omega}\mathbf{B}^T$ *and* $\mathbf{S} = \mathbf{\Theta}\mathbf{A}^T$. *Then it holds:*

$$\mathbf{S} = \mathbf{\Theta}\mathbf{C}\mathbf{B}^T, \tag{6}$$

where $\mathbf{C} = \mathbf{\Theta}^T\mathbf{\Omega} = \mathbf{A}^T\mathbf{B}$ *is a* $K_t \times K_s$ *matrix of spatiotemporal coefficients.*

Above theorem is a case of a perfect reconstruction. It is also possible to consider a reduced base model, where the number of basis vectors is significantly smaller than the original number. Furthermore, it is worth to note that if $K_s \ll 3P$ and $K_t \ll F$ than the $K_t \times K_s$ coefficients in $\mathbf{C}$ can be orders of magnitude fewer than the $F \times K_s$ coefficients in Ω or the $K_t \times 3P$ coefficients in $\mathbf{A}$. The following theorem [3] gives us an estimation of a reconstruction error of the bilinear spatiotemporal model in such a case. The $||\cdot||_F$ is the Frobenius norm.

Theorem 2. *Let* $\epsilon_t = ||\mathbf{S} - \mathbf{\Theta}\mathbf{A}^T||_F$ *is the reconstruction error of the trajectory model and* $\epsilon_s = ||\mathbf{S} - \mathbf{\Omega}\mathbf{B}^T||_F$ *is the reconstruction error of the shape model. Then for the reconstruction error of the bilinear spatiotemporal model* $\epsilon = ||\mathbf{S} - \mathbf{\Theta}\mathbf{C}\mathbf{B}^T||_F$ *we have* $\epsilon \leq \epsilon_t + \epsilon_s$.

For a shape basis $\mathbf{B}$ and a trajectory basis $\mathbf{\Theta}$, we compute the bilinear model coefficients $\mathbf{C}$, minimizing the reconstruction error for a given $\mathbf{S}$. We will use formula appropriate for the orthonormal bases [3]:

$$\mathbf{C} = \mathbf{\Theta}^T\mathbf{S}\mathbf{B}. \tag{7}$$

2.2 Considered orthonormal bases

In the original paper [3] proposing the bilinear spatiotemporal model, there were suggested two orthonormal bases - PCA and DCT based. The choice is very reasonable. The PCA is capable to adopt to virtually any set of body pose configurations or trajectories. The PCA can be obtained [9] through singular value decomposition:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \tag{8}$$

hence the transform is given as:

$$\mathbf{T} = \mathbf{X}\,\mathbf{V}, \tag{9}$$

where: $\mathbf{V}$ contains the eigenvectors as columns, $\mathbf{\Sigma}$ nonzero singular values.

Alas, for long sequences, obtaining full trajectory PCA can be computationally intensive and may require a relatively large amount of memory (gigabytes). The natural alternative is the commonly used base of Discrete Cosine Transform (DCT). The DCT is good approximation of PCA for 'natural' signals - also for motion - so trajectories should be represented well. On the other hand, it is demonstrated in the experimental part of the paper, one should not expect good performance in the shape representation. The DCT base [1] is given as:

$$D_{u,f} = \begin{cases} \dfrac{1}{\sqrt{F}} & \text{, for } u = 0,\ 0 \leq f \leq F-1 \\ \sqrt{\dfrac{2}{F}} \cos \dfrac{\pi(2f+1)u}{2F} & \text{, for } 0 \leq u \leq F-1,\ 0 \leq f \leq F-1 \end{cases} \tag{10}$$

3 Experiments

In order to reveal efficiency of the bilinear model we performed two experiments. First, to obtain overview of the efficiency in a function of K_s and K_t we performed exhaustive evaluation of model accuracy with RMSE. Each of the basis combination was tested. These results allowed us to neglect two of the combinations and further experiments on the selection appropriate K_s and K_t for the fixed number of coefficients were performed using selected approaches only.

3.1 The data

For the testing purposes, we selected small, yet comprehensive set of MoCap recordings. It contains sequences recorded using different parameters: speed (60-200Hz), lengths and number of markers (25-53) of various subjects and actions. Two subjects - male HJ and female IM - *Range of movement* (ROM) sequences (exercising all limbs and all rotation extremes for every joint) which caused the large variance in poses. Ordinary motion of human and non-human (dog) subjects that have a smaller variance in poses. Two non-rigid facial animations demonstrating simple spelling of the alphabet and a kind of 'ROM' for facial expressions (with head movements) - presenting smaller and larger pose variance respectively. Finally a hands typing the keyboard sequence which has limited variance in poses.

Table 1. Experimental MoCap sequences

No	Name	Description of sequence	frames (F)	mark. (P)
a)	HJ-rom	range of movement a male subject	10486 (52.4sec@200Hz)	53
b)	IM-rom	range of movement a female subject	3675 (36.7sec@100Hz)	53
c)	HJwalk	walk - turn (180deg) - walk	1857 (9.3sec@200Hz)	53
d)	HJsit	Tpose-sit-standup	1618 (8.1sec@200Hz)	52 (1 lost)
e)	Dog[1]	dog run, jump, turn, walk, step onto and off the table	717 (11.9sec@60Hz)	25
f)	Face-exp	head moves, ROM for expressions emphasis on mouth and eyebrows	3918 (65.3sec@120Hz)	45
g)	Face-say	face spelling alphabet	1780 (17.8sec@100Hz)	36
h)	Hands	Both hands typing the keyboard	794 (7.9sec@100Hz)	40

3.2 Overview of the efficiency of bases

In the experiment we examined all four combinations of PCA and DCT as spatiotemporal basis. The evaluations with the RMSE were performed exhaustively for a number of base vectors varying between 1 and 20. It provided a general overview of the performance of each type of basis and a brief view into the performance gain with the growing number of coefficients. The interpretation of these can be performed on observation how fast an error reduces with the growing number of base vectors. If it reduces slowly for the PCA basis, it implies more varied movement (larger number of poses). When we observe relatively large

[1] Sequence from 'Dog package' (`http://www.mocapclub.com/Pages/Library.htm`)

error (slow reduction) up to the large number of DCT base vectors, it suggests the presence of high frequencies and therefore fast motions in the sequence.

In the Fig. 1 we see the results for a ROM sequence for HJ subject, the results for IM were almost the same they are not illustrated. Wide set and range of possible body poses cause slower error decreasing with increasing number of base functions than it is in case of more ordinary motion sequences such as walk.

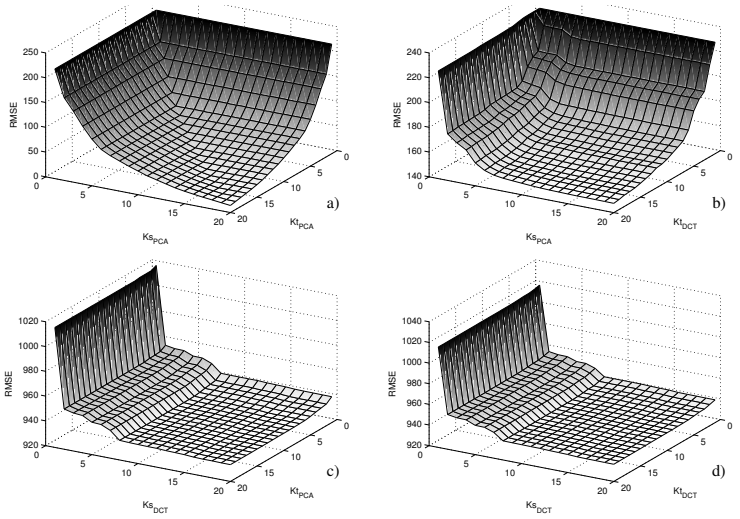

Fig. 1. Comparison of all combinations of spatio-temporal bases for ROM of HJ subject: a) PCA-PCA, b) PCA-DCT c) DCT-PCA d) DCT-DCT. (Please mind the scales)

Ordinary motions for sit, walk and dog sequences also share a similar characteristics in bilinear model. The Fig 2 illustrates a representative example (walk sequence). Such sequences demonstrate limited variability in the poses - error reduces very fast with the growing number of PCA base vectors. Also for the DCT in temporal domain error decays fast as the number of base vectors increases.

Facial and hand sequences shared another, common characteristics - please see the Fig. 3 as representative one. They demonstrate quite a limited set of poses (especially spelling face and hands) so to represent their shape one needs relatively small amount of base vectors. Although, in the temporal domain of these cases, we observe that error reduces relatively slowly with the growing amount of DCT base vectors, so one can suspect that there are fast motions (frequencies) present in these recordings.

The test revealed the best performance of PCA-PCA basis with diagonal-symmetric RMSE reduction with the growing number of coefficients. The DCT appeared to perform poorly as a base for the representation of complex shape structure, whereas it is a good representation for the temporal data. So in further tests we rejected both model combinations based on the DCT as a shape basis.

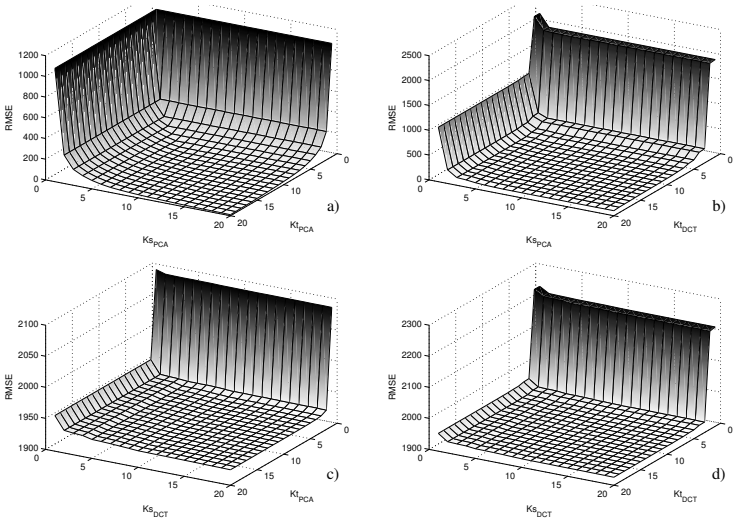

Fig. 2. Comparison of all combinations of spatio-temporal bases for HJ walk: a) PCA-PCA, b) PCA-DCT c) DCT-PCA d) DCT-DCT. (Please mind the scales)

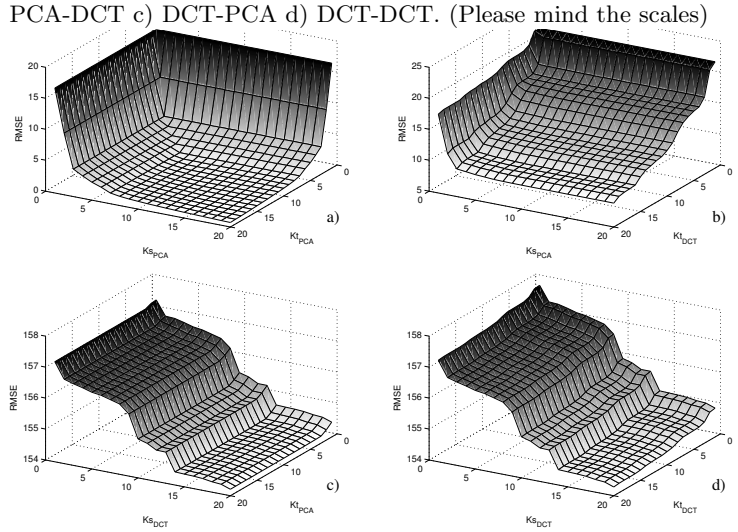

Fig. 3. Comparison of all combinations of spatio-temporal bases for facial expressions: a) PCA-PCA, b) PCA-DCT c) DCT-PCA d) DCT-DCT. (Please mind the scales)

3.3 Optimal division of base vectors number

In the case when we have a limited number of spatiotemporal coefficients - fixed N - there is a problem how to divide N between K_s and K_t. To address this question we conducted some dedicated tests, which were a series of RMSE evaluations of bilinear model for a number of feasible divisions. Since the coefficients array is rectangular where $N = K_s \cdot K_t$ it is not possible to get integer sizes in every case. Therefore, we decided to check the RMSE for N which was constant only approximately and so the characteristics in Fig. 5 might look a bit 'jaggy'. The iterations were for $K_s^i = K_s^{\min}, \ldots, K_s^{\max}$ where:

$$K_s^{\min} = \left\{ \begin{array}{ll} \text{if } N > F & : \lceil N/F \rceil \\ \text{else} & : \quad 1 \end{array} \right\}, \qquad K_s^{\max} = \left\{ \begin{array}{ll} \text{if } N > 3P & : \ 3P \\ \text{else} & : \ N \end{array} \right\},$$

and K_t was evaluated as $K_t^i = \text{round}\left(\frac{N}{K_s^i}\right)$.

Let's define $\alpha = K_s/(K_s + K_t)$ a relative share of spatial bases in the overall number of base vectors. We are looking for the α^{opt} resulting in minimal error. Having chosen α it is easy to show the numbers of base vector numbers are:

$$K_s = \text{round}(\sqrt{\frac{N\alpha}{1-\alpha}}), \qquad K_t = \text{round}(\frac{N}{K_s}). \tag{11}$$

Due to neglecting (in p. 3.2) of the usability of DCT as a base for the structured information (shape), further tests were performed using PCA-PCA and PCA-DCT base combinations only. The results are presented in Fig. 5 for all the test datasets. The reconstruction error for each of the datasets was tested, against various overall numbers of coefficients: $N = 2^2,\ 2^5,\ 2^{10},\ 2^{15},\ 2^{20},\ 2^{30},\ 2^{50}$.

The PCA-PCA base characteristics, with the optimal equal division of base vectors($\alpha^{opt} = 0.5$), is obvious because of the symmetry of importance of shape/time bases observed in p. 3.2. Although, the PCA-DCT result is not so trivial. We found out that $\alpha^{opt} \in (0.1,\ 0.3)$ so the shape base vectors should be approximately 10-30% of an overall number (see dashed lines in Fig. 5). As one can note there is a single and well visible minimum in the characteristics, so other choices of dimensions result in degradation of reconstruction. Such observation is consistent for all the test cases so 20-80 division might be useful suboptimal choice (see Fig. 4) when we cannot search for optimal division.

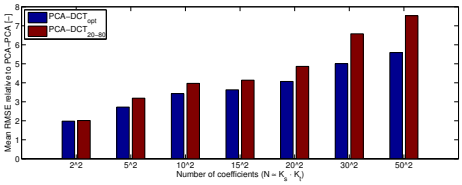

Fig. 4. Mean RMSE of PCA-DCT relative to PCA-PCA for α_{opt} and 20-80 setup

The visual demonstration of results obtained is presented in the Fig. 6. It includes original body poses and reconstructions from bilinear representation with both considered bases for various and (sub)optimal base sizes. We can observe that PCA-DCT gradually (and slower) converges to the original shape with the growing number of coefficients, whereas PCA-PCA reaches proper shape quite fast and the shape is just a bit refined with the larger number of coefficients.

4 Summary

The bilinear representation for a motion capture sequences is a novel idea. In this study, we analyzed its performance using combinations of two fundamental basis (PCA and DCT) for their reconstruction error. We verified commonly known fact that there is no use to employ the DCT as a shape basis. The most

obvious result is for PCA-PCA basis. It is symmetric and results in the best error reduction with the growing number of coefficients, which result in 50-50 optimal sharing between shape and trajectory.

The most notable result, we obtained for the PCA-DCT as a basis. We discovered interesting property for division of a base vectors number between spatial and temporal parts for respectively PCA and DCT bases. The optimal division favors trajectory bases with the relative share of shape bases in the overall number between 0.1 and 0.3, therefore, one could consider 20-80 division.

Further research will focus on using the bilinear model for pattern recognition. In such a case it would be necessary to use the ability of class discrimination as performance assessment criteria. Also, other basis should be also taken into consideration instead of PCA - ICA or LDA seem to be appropriate.

Acknowledgments. This research has been supported by Demonstrator+ Programme of NCRD. Project UOD-DEM-1-183/001 "System inteligentnej analizy wideo do rozpoznawania zachowa i sytuacji w sieciach monitoringu".

References

1. Ahmed, N., Natarajan, T., Rao, K.: Discrete Cosine Transform. IEEE Transactions on Computers C-23(1), 90–93 (Jan 1974)
2. Akhter, I., Sheikh, Y., Khan, S., Kanade, T.: Nonrigid Structure from Motion in Trajectory Space. In: Koller, D.e.a. (ed.) Adv. in Neural Inf. Proc. Sys. 21 (NIPS2008), pp. 41–48 (2009)
3. Akhter, I., Simon, T., Khan, S., Matthews, I., Sheikh, Y.: Bilinear spatiotemporal basis models. ACM Transactions on Graphics 31(2), 1–12 (Apr 2012)
4. Bregler, C., Hertzmann, A., Biermann, H.: Recovering non-rigid 3d shape from image streams. In: IEEE Conf. on Comp. Vis. and Patt. Rec. pp. 690–696 (2000)
5. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active Shape Models;Their Training and Application. Comput. Vis. Image Underst. 61(1), 38–59 (Jan 1995)
6. Lucey, P., Bialkowski, A., Carr, P., Morgan, S., Matthews, I., Sheikh, Y.: Representing and Discovering Adversarial Team Behaviors Using Player Roles. In: 2013 IEEE Conf. on Comp. Vis. and Patt. Rec. (CVPR). pp. 2706–2713 (Jun 2013)
7. Moeslund, T.B., Granum, E.: A Survey of Computer Vision-Based Human Motion Capture. Comput. Vis. Image Underst. 81(3), 231–268 (Mar 2001)
8. Moeslund, T.B., Hilton, A., Krueger, V.: A Survey of Advances in Vision-based Human Motion Capture and Analysis. Comput. Vis. Image Underst. 104(2), 90–126 (Nov 2006)
9. Shlens, J.: A tutorial on principal component analysis. arXiv, Computing Research Repository (CoRR) abs/1404.1100 (2014), `http://arxiv.org/abs/1404.1100`
10. Skurowski, P., Pawlyta, M.: Functional body mesh representation - a simplified kinematic model. In: AIP Conf. Proc. vol. 1648, pp. 660008–1–4 (Mar 2015)
11. Torresani, L., Bregler, C.: Space-Time Tracking. In: Proc. of the 7th European Conf. on Comp. Vis. (ECCV). LNCS, vol. 2351, pp. 801–812. London (2002)
12. Wei, X., Sha, L., Lucey, P., Morgan, S., Sridharan, S.: Large-Scale Analysis of Formations in Soccer. In: 2013 Int. Conf. on Digital Image Comp.: Techn. and Appl. (DICTA). pp. 1–8 (Nov 2013)

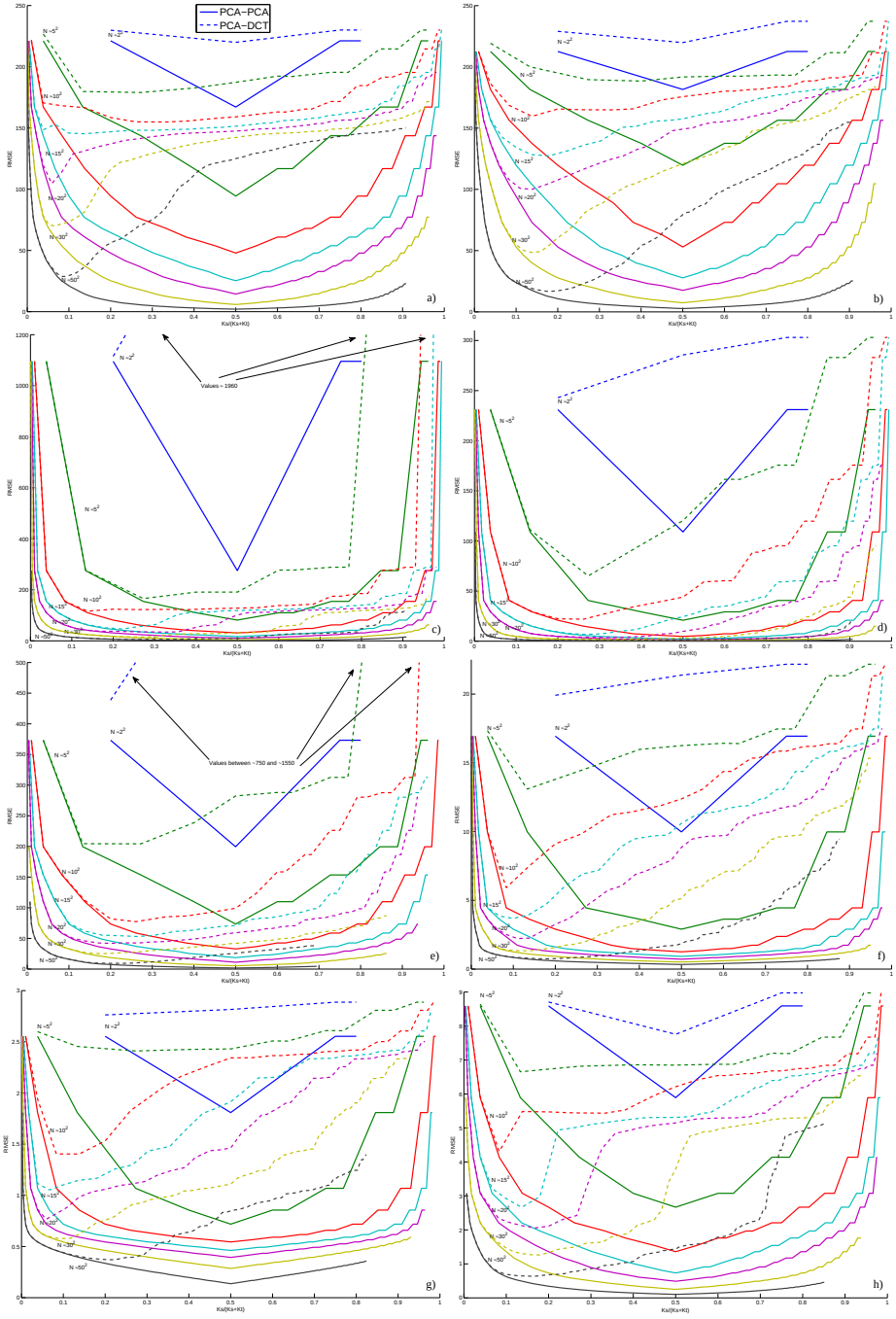

Fig. 5. Evaluation of bilinear model for various $\alpha = K_s/(K_s + K_t)$ coefficients division for the PCA-PCA (—) and PCA-DCT (- -) bases - source sequences (a-h) as in Tab. 1

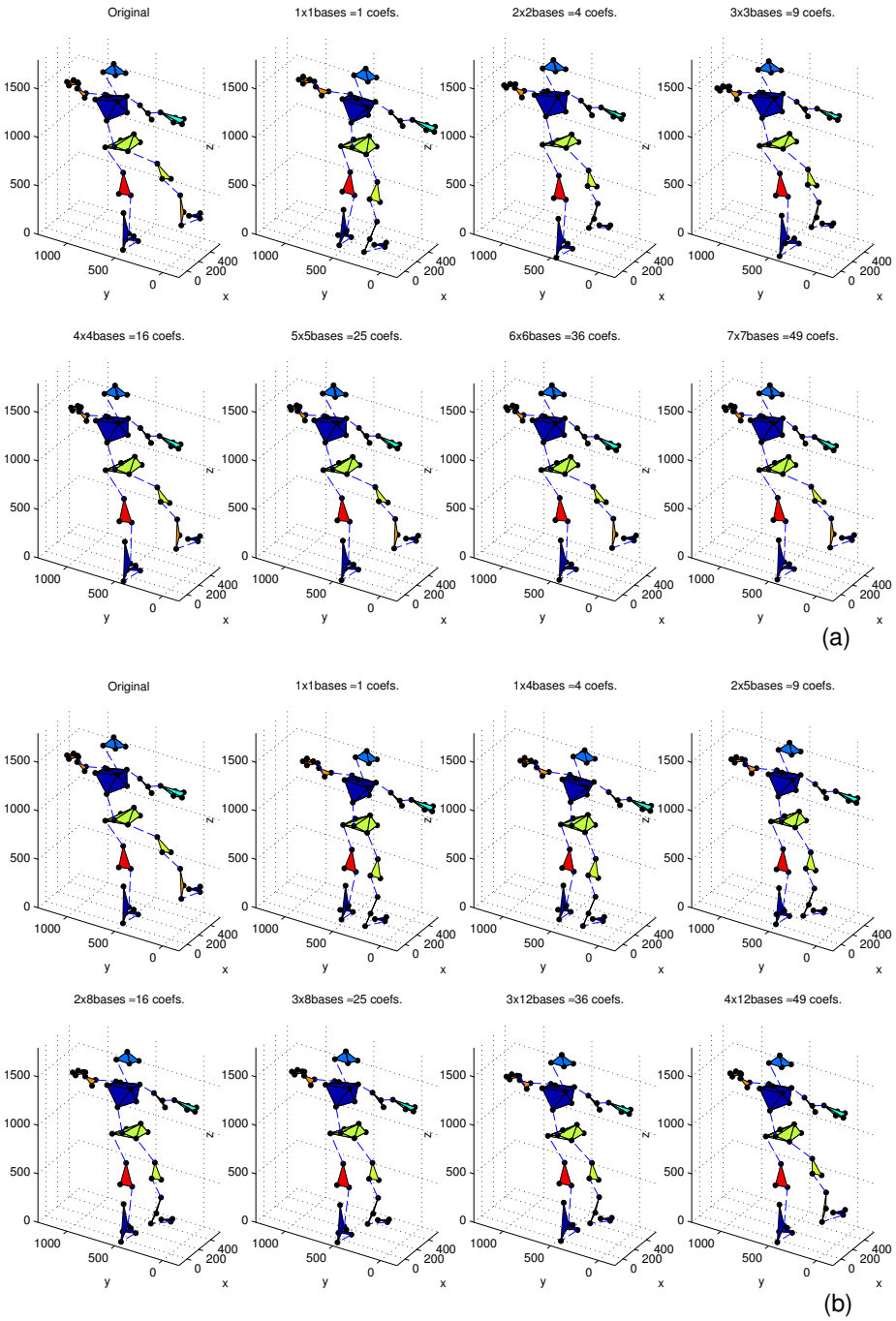

Fig. 6. A single frame from HJrom sequence reconstructed with increasing number of PCA-PCA (a) and PCA-DCT (b) bases (20-80) - body visualized with FBM [10]