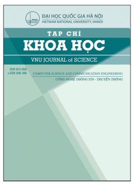
VNU Journal of Science: Comp. Science & Com. Eng., Vol. 35, No. 1 (2019) 1–10

VNU Journal of Science:
Computer Science and Communication Engineering

Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce



Learning and transferring motion style using Sparse PCA

Do Khac Phong[1*], Nguyen Xuan Thanh[1], Hongchuan Yu[2,]

[1]*Faculty of Information Technology, VNU University of Engineering and Technology,*
No. 144 Xuan Thuy Street, Dich Vong Ward, Cau Giay District, Hanoi, Vietnam
[2]*National Centre for Computer Animation, Bournemouth University,*
Talbot Campus, Fern Barrow, Poole, Dorset, BH12 5BB, United Kingdom

Abstract

Motion style transfer is a primary problem in computer animation, allowing us to convert the motion of an actor to that of another one. Myriads approaches have been developed to perform this task, however, the majority of them are data-driven, which require a large dataset and a time-consuming period for training a model in order to achieve good results. In contrast, we propose a novel method applied successfully for this task in a small dataset. This exploits Sparse PCA to decompose original motions into smaller components which are learned with particular constraints. The synthesized results are highly precise and smooth motions with its emotion as shown in our experiments.

Received 07 May 2018, Revised 03 December 2018, Accepted 29 December 2018

Keywords: Sparse PCA, style learning, motion style transfer

1. Introduction

The automatically precise stylization of human motion to express mood is a primary role in realistic humanoid animation. Motion style transfer is a primary problem in computer animation, allowing us to convert the motion of an actor to that of another character. Such characters can express their emotions like happy, sad, joy, or so. The precise stylization of human motion to express the state of mind or identity plays a vital role in realistic humanoid animation. Previously, this manual work takes numerous time to generate huge variations of motion data, thereby automating this process is really useful and essential for a bunch of applications such as films, and computer games.

Many approaches have been developed

[*] Corresponding author. Email: phongdk@vnu.edu.vn

https://doi.org/10.25073/2588-1086/vnucsce.206

for this style transfer task. Hsu et al. [1] introduced a linear time-invariant (LTI) model for homogeneous human motion stylization, e.g. walking. For heterogeneous motion, Xia et al. [2] proposed a method through temporally local nearest neighbor blending in spatial-temporal space. Recently, along with the rapid exploration of deep learning technique, neural style transfer for images is introduced by Gatys and his colleagues [3]. Inspired by their work, Holden et al. [4] adapted it and used a deep neural network to transform a style of motion data. Nevertheless, these data-driven methods require a large number of training datasets and manual alignment leading to typically time-consuming.

In this paper, our goal is to build a framework for the rapid style transfer process, as well as eliminating the training time. To do this, we first decompose original motions into smaller factors: weights and components. A style transfer algorithm is then performed to generate transferred components following by a motion reconstruction step that satisfies the constraints with respect to content, style and bone length of a target character.

To summary, our main contributions are:

(a) Propose a novel, fast and effective model to transfer motion based on matrix factorization.

(b) Our model can be applied to small datasets.

2. Related Work

2.1. Matrix factorization

Matrix factorization is a technique that factorizes a single matrix into a product of matrices. This could be understood as a way to find a new representation of data with much lower dimensions or to dimension reduction. In particular, Principle Components Analysis (PCA) is a primary and popular method that decompose multivariate data into a set of orthogonal components. In other words, PCA attempts to represent each principal component by a linear combination of the original variables such that the derived variables capture maximal variance [5]. Nevertheless, the coefficients of all variables are typically nonzero causing a difficulty in the derived principal components interpretation. It is obvious that the global effects are not essential in some circumstances. For example in face decomposition, sparse components extracted should be an eye, a nose.

Sparse PCA, in contrast, is a variant of PCA which produce localized components [6], [5] by introducing a sparsity-inducing norm such as l_1. Such methods exploited a localized set of variables, thereby applying successfully in computer vision, medical imaging and signal processing. Neumann et al. [7] extended Sparse PCA for animation processing by adding local support map which is suitable for surface deformations, for instance, faces or muscle. Localized components are appropriate for motion-emotion data where the state of mind is shown via actions, and each action is associated with several human parts. Such work inspired us to use Sparse PCA in this paper.

2.2. Correlation, Covariance and Gram matrix

Suppose we have a set of centered column vectors $X_i \in R^{m\times 1}$, $i = 1, \dots, n$; forming a matrix $X = [X_1, X_2, \dots, X_n]$, $X \in R^{m\times n}$, and $m > n$. A Singular Value Decomposition (SVD) of X expresses it as $X = UDV^T$ where $D_{k\times k}$ is an diagonal matrix with positive values which are the "singular values" of X on the diagonal, $U_{m\times k}$ and $V_{k\times n}$ are unitary matrices.

The covariance and correlation matrix of X, denoted as Σ_X and r_X respectively are computed by the following formulas:

$$\Sigma_X = E[XX^T] = \frac{1}{m-1}XX^T = UD^2U^T \quad (1)$$

$$r_X = (diag(\Sigma_X))^{-1/2}\Sigma_X(diag(\Sigma_X))^{-1/2} \quad (2)$$

where $diag(\Sigma_X)$ is the matrix of the diagonal elements of Σ_X. The correlation matrix can be seen as the covariance matrix of the standardized X_i. Meanwhile, the Gram matrix of X, denoted as $Gram(X)$, is calculated as:

$$Gram(X) = X^TX = VD^2V^T \quad (3)$$

As can be seen from Eq.1 and Eq. 3, the gram matrix and the covariance matrix share the same eigenvalues up to the $(m-1)$ factors. Therefore, minimizing the difference of two matrices using their covariance or correlation matrices is equivalent to optimizing their Gram matrices. That is reason why many techniques, e.g. Multi-Dimensional Scaling, Kernel PCA use Gram matrix to compute the principal components instead of covariance matrix [8], [9] in case of $m \gg n$. Additionally, Gatys et al .[3] exploited Gram matrix to calculate the features correlation in style representation of an image towards transferring its style to others.

2.3. Motion Style transfer

Basically, human motion expresses the action it embodies whereas a considerable component of a natural human act is the style of that action. Furthermore, the style and emotion of a motion are more likely to convey meaningful information compared with the underlying motion itself. The accurate stylization of human motion to express mood or identity is a key role in realistic humanoid animation. This works benefits a wide range of applications, especially virtual games, films instead of capturing an enormous amount of all possible actions and styles [2],[10].

Many solutions have been proposed for human motion stylization and most of them are data-driven techniques. A linear time-invariant model was proposed by [1] to encode style differences between motions, and the learned model was utilized to transfer motion from one style to another style in a real time. However, this method was developed for homogeneous human behaviors, e.g. walking, kicking. Xia et al. [2] introduced a series of a local mixture of autoregressive models to capture complex relationships between styles of heterogeneous motions (walking ⇒ running ⇒ jumping), and a search scheme to seek appropriate style candidate on a huge motion dataset.

Yumer et al. [10] developed a method where the style transfer task is performed in the frequency domain. Nevertheless, their method requires a costly searching step to find the best candidates from an available database for a source style and reference style in the spectral domain. Holden et al. [4] applied convolutional autoencoder network to perform the style transfer task over the neural network hidden unit values to generate a motion that has the content of one input but with the style of another. A large motion database collected from many different sources of motion capture (CMU[1], Xia et al.[2], etc), was converted into a suitable format for training the neural network that is a time-consuming process. Such work promotes us to discover a new strategy which can apply for a small stylized motion data set.

3. Methodology

The architecture of our model to transfer the style of a motion M_C to a target (content) motions M_S is shown in Fig. 1. Each motion $M \in R^{F\times 3N}$ consists of a mesh animation with F frames in which each frame f is a pose with N joint positions in 3D. Then, the corresponding components C_S (hidden style representation) and C_C (hidden content representation) of two input motions are extracted in the decomposition step. In style transfer process, a white noise component C_X is adjusted such that it matches both components C_S and C_C. Finally, the new motion M_X is composed of the mixed components $\tilde{C}_X$ and the target weights W_C.

3.1. Decomposition

Given a motion $M \in R^{F\times 3N}$, we seek an appropriate matrix factorization technique to decompose M into K deformation components C with weights W

$$M_{F\times 3N} = W_{F\times K}.C_{K\times 3N} \tag{4}$$

The matrix W with the one dimension F is assumed to include time variant data, meanwhile the matrix C contains coordinates of K basic motions (see section 3.3 for more details). Depending on the regularization term added to Eq.(4), there are many different solutions for W and C. In PCA, this constraint is the orthogonality of the components, $C^T C = I$. On the other hand, by imposing sparsity reducing norm such as l_1 norm, sparse components can be achieved in *Sparse PCA* [5]. Subsequently, the matrix factorization now turns into a joint regularized minimization problem as:

$$\underset{W,C}{\arg\min} \; ||M - W.C||_F^2 + \Omega(C) \qquad s.t.\phi(W) \tag{5}$$

Since the i^{th} joint in frame k is identified by a triplet coordinate $j_k^{(i)} = [x, y, z]_k^{(i)}$, while regularizing C with l_1 norm could induce sparsity, the group structure would be ignored leading to eliminating each dimension separately. Consequently, the $l1/l2$ norm is utilized to make the dimension vanish simultaneously [11],[7], [12]

$$\Omega(C) = \sum_{k=1}^{K} \sum_{i=1}^{N} ||j_k^{(i)}||_2 \tag{6}$$

The direct optimization of Eq.(5) is

[1] http://mocap.cs.cmu.edu/

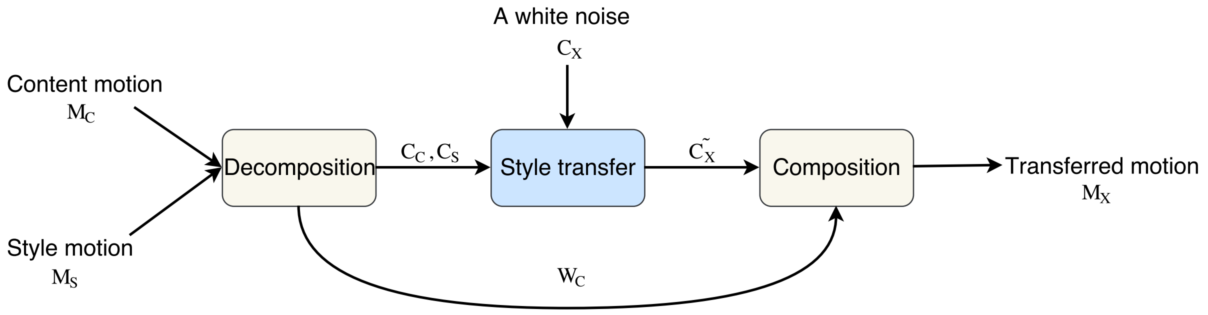


Figure 1. Our framework for motion style transfer.

complicated due to its non-convex. By fixing either W or C, the problem is convex and can be solved easily by an iterative refinement method that alternates between the two optimization tasks [13], [7].

3.2. Style Learning

Our idea in order to learn a new style for the target motion is transforming basic motions C_C to C_S resulting in stylized components $\bar{C_X}$. In other words, C_X is matched with both the content representation and style representation.

3.2.1. Content

As expected the transferred motion contains the content of the target motion M_C, the difference between the content representation of the target motion and of the transferred one is considered as the content loss of our model:

$$L_{content} = c||C_C - C_X||^2 \tag{7}$$

where the user-defined scaled weight c is set to 1.0 in our experiments.

3.2.2. Style

In order to transfer the style of the input motion M_S to the content motion M_C, the style loss is defined as the distinction between the style representation of the input style motion and of the transferred one. This is scaled by a user-specified weight s ($s = 0.01$ in our cases) as follows:

$$L_{style} = s||Gram(C_S) - Gram(C_X)||^2 \tag{8}$$

where the Gram matrix calculate the inner products of the element values in components C across basic motions

$$Gram(C) = \sum_i^K C_i^T C_i \tag{9}$$

3.2.3. Constraint

Although we are able to achieve a stylized motion by a multiplication of the content weights W_C and the stylized components C_X by optimizing Eq. (7) and (8), it does not guarantees the composed stylized motion in a human body form. As a consequence, additional loss function related to human bone length is exploited as a constraint to human body. Suppose that each bone b in transformed motion M_X has two joints j_1 and

j_2, so their bone length is a distance in 3D space of their coordinate $p_{b_{j_1}}$ and $p_{b_{j_2}}$. Given a length l_b, the bone length constraint is in a form:

$$L_{bone} = \sum_b |||p_{b_{j_1}}^{M_X} - p_{b_{j_2}}^{M_X}|| - l_b|^2 \qquad (10)$$

The stylized components C_X first is initialized from white noise. Afterwards, it is adjusted via stochastic gradient descent with automatic derivatives calculation performed via Theano until the following total loss converges to a particular threshold

$$L_{total} = L_{content} + L_{style} + L_{bone} \qquad (11)$$

To speed up the process learning the stylized components, Adam [14] is used for stochastic optimization in our experiment.

3.3. Composition

Since the original motion is decomposed into the K basic motions, the synthesized motion is an inverse process indeed. The third-row figure in Fig. 2 shows the reconstruction motion utilizing the first two basic motions (first two rows of the matrix C with $K = 30$), which is able to approximate 70% the content of the original one. For the first four basic motions (the bottom figure), the majority of the content motion is preserved in spite of not being too smooth as the origin. Our purpose is to retain the personality of the content motion, so the target weight matrix W_C is kept unchanged and taken as input of the composition process, along with the transferred components C_X output from the style transfer step. Simply, a transferred motion M_X is reconstructed in a form:

$$M_X = W_C.\tilde{C}_X \qquad (12)$$

4. Experimental Results

4.1. Dataset

We collect freely available Emotional Body Motion Database[2] which consists of 1447 files in BVH format [15], [16]. Notwithstanding, we only keep 323 files which have an agreement between two fields: ‘Intended emotion’ and ‘Perceived category’. The former represents the emotion the actor intended to convey, whereas the latter shows the emotion was chosen by most of the observers. There are eight participants (four males and four women) freely showing 11 emotion categories namely amusement, anger, disgust, fear, joy, neutral, pride, relief, sadness, shame, and surprise via their entire body, face, and voice. Nevertheless, our study focuses on their skeleton motion only. Additionally, the emotion is expressed mostly by their upper body movement as we observed.

All of the motion in the database are downsampled to 60 frames per second (fps) from 120 fps and converted into the 3D joint position format from rotational motion in the original dataset. The origin is on the ground where the root position is projected onto. In addition, the joint positions are located in the body’s local coordinate system. Finally, we subtract mean pose from data, then divide by their own standard deviation resulting in zero mean and standard deviation motion

[2] http://ebmdb.tuebingen.mpg.de/

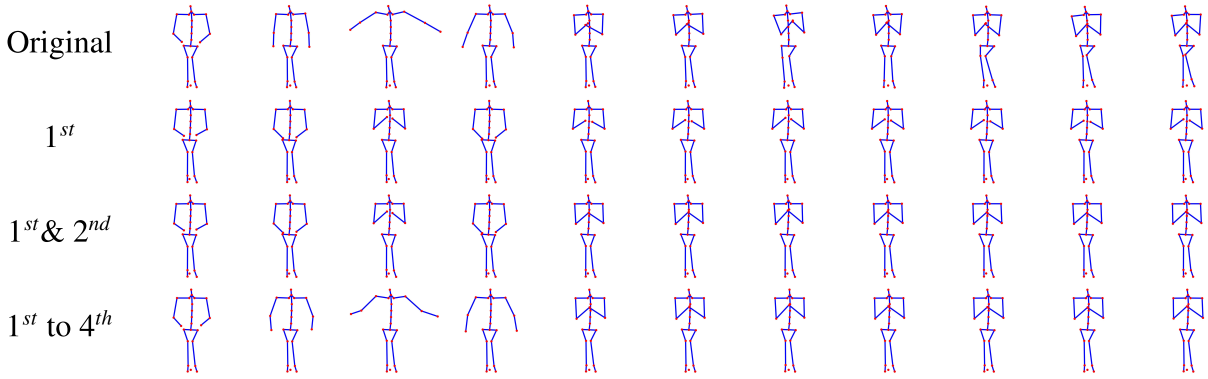


Figure 2. Reconstruction using several basic motions (K=30)

data. Each pose is represented by the 23 joint positions giving us 69 degrees of freedom (DOF) in total. Although the motion duration can be either different or fixed, each motion in our experiment has similar length, and last for about 200 frames.

4.2. Results

4.2.1. Stylization

In this section, we demonstrate some results of our approach. As can be seen from Fig. 3, the first character action describe *Pride* mood whilst the behavior of the second one is *Disgust*. The transferred motion using Sparse PCA retains the personality of the former, but with the latter's style. Consequently, we achieve a new motion in *Disgust* mood. Besides, we also take into account the effect of parameter K in our experiment. For $K = 10$, the left-hand folds too tight and it looks less similar to the input style figure than those with $K = 30$ or $K = 50$. Meanwhile, the spine in some frames for $K = 50$ is not straight as in the corresponding frames for $K = 30$. This suggests that if we retain a number of deformation components too few, it will lose more information and make the transferred motion less natural. The similar outcome is indicated in Fig. 4 where the new motion is synthesized from two motions in *Surprise* and *Anger* mood. The stylized motion in our model behaves in a way that he/she is *Anger* and the most similar one is when $K = 30$.

4.2.2. Sparsity

Fig.3 and Fig. 4 demonstrates the effects of sparse decomposition as well. In spite of learning style components of PCA, the style of synthesized motion is not transferred precisely as contrast to the stylized motion in our method using Sparse PCA. It indicates that localized components are better than global components in animation. In addition, the group structure controlled by Eq.6 also makes the dimension be modified simultaneously, contributing to spatial preservation.

4.2.3. Constraint

The advantageous of the bone length constraint is demonstrated explicitly in Fig. 4. In a case of missing L_{bone}, the left hand of the synthesized motion is longer than that of the *Surprise* one. In contrast, the

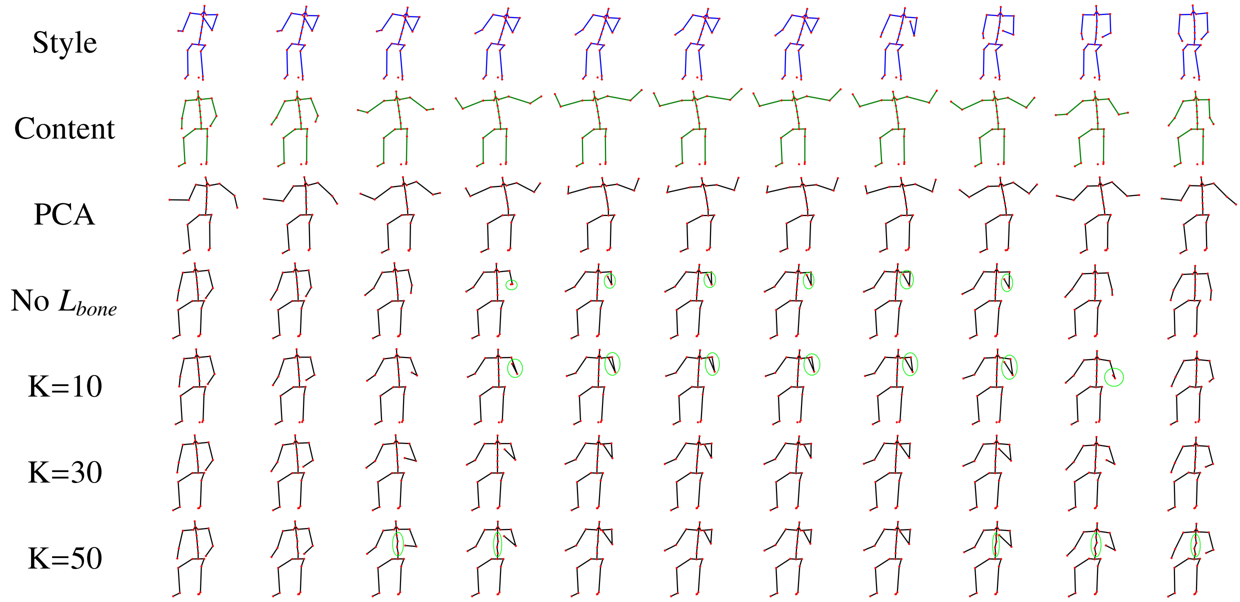


Figure 3. Animations are generated in time series. Blue: input style motion (Disgust). Green: input content motion (Pride). Black: output transferred motion. Green circles/ellipses are invalid shapes. The last four row used Sparse PCA.

shorter left hand is indicated in Fig. 3, compared to the target motion in *Pride* mood. The explanation for this is that during the iterative period of learning components and reconstructing motion, the difference between the target and synthesized motion with respect to bone length is minimized, finally making the stylized motion capture the human body form of the target one.

5. Conclusions

In this paper, we introduce a novel algorithm for motion style transfer task. Our method is an integration of matrix factorization and an artistic style learning technique. This work can deal with a shortage of large motion datasets since it can be applied to small ones. In spite of gaining some promising outcomes, there are several limitations remaining:

1. The number of deformation components (K) has to be defined in advance.

2. The tuning parameters s and c are user-specified. It is a trade-off between the style and content we want to transfer to a new motion.

3. The style representation between two motions is minimized during the learning period, and in fact the best style is unknown.

4. The velocity and acceleration of the human body are omitted in this study which are vital properties to make motion smooth and natural.

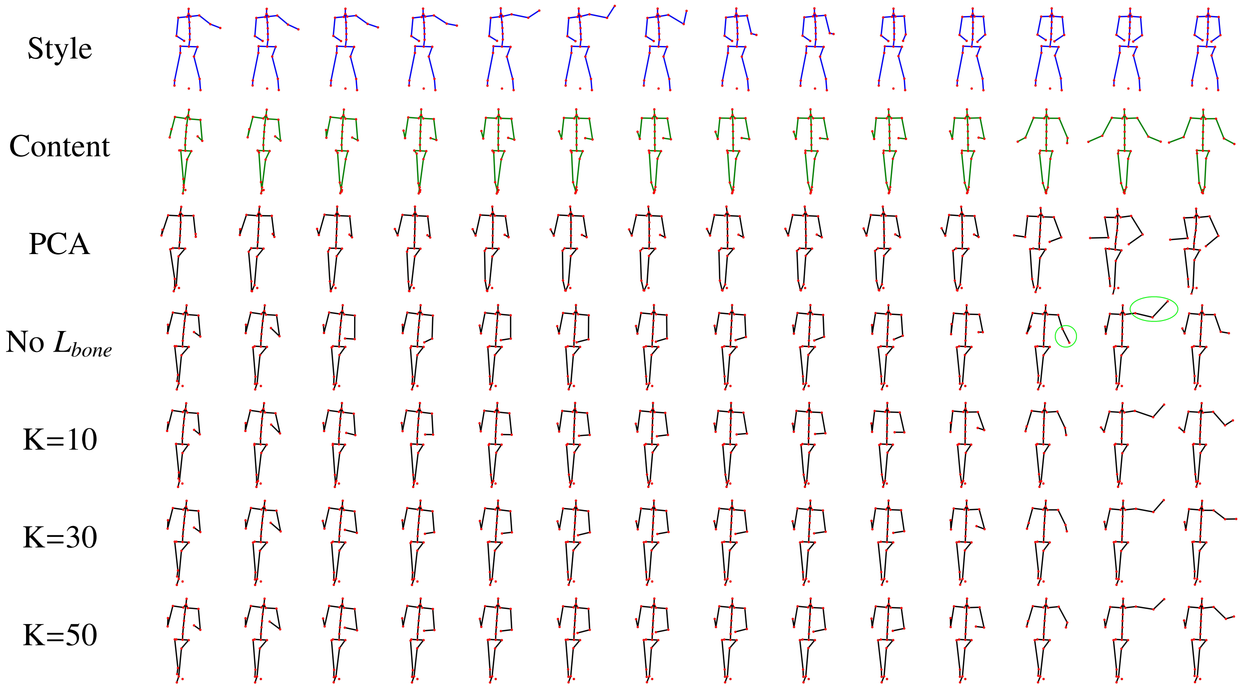


Figure 4. Animations are generated in time series. Blue: input style motion (Anger). Green: input content motion (Surprise). Black: output transferred motion. Green circles/ellipses are invalid shapes. The last four row used Sparse PCA.

Those limitations are perceived as our future work.

Acknowledgments

We would like to thank anonymous reviewers for their detailed comments on our paper. This work was supported by EU H2020 project-AniAge (No.691215), and by the project named “Multimedia application tools for intangible cultural heritage conservation and promotion” (No. DTDL.CN-34/16). The emotional body motion database was provided by the Max-Planck Institute for Biological Cybernetics in Tuebingen, Germany.

References

[1] E. Hsu, K. Pulli, J. Popović, Style translation for human motion, in: ACM Transactions on Graphics (TOG), Vol. 24, ACM, 2005, pp. 1082–1089.

[2] S. Xia, C. Wang, J. Chai, J. Hodgins, Realtime style transfer for unlabeled heterogeneous human motion, ACM Transactions on Graphics (TOG) 34 (4) (2015) 119.

[3] L. A. Gatys, A. S. Ecker, M. Bethge, A neural algorithm of artistic style, arXiv preprint arXiv:1508.06576.

[4] D. Holden, J. Saito, T. Komura, A deep learning framework for character motion synthesis and editing, ACM Transactions on Graphics (TOG) 35 (4) (2016) 138.

[5] H. Zou, T. Hastie, R. Tibshirani, Sparse principal component analysis, Journal of computational

and graphical statistics 15 (2) (2006) 265–286.

[6] I. T. Jolliffe, N. T. Trendafilov, M. Uddin, A modified principal component technique based on the lasso, Journal of computational and Graphical Statistics 12 (3) (2003) 531–547.

[7] T. Neumann, K. Varanasi, S. Wenger, M. Wacker, M. Magnor, C. Theobalt, Sparse localized deformation components, ACM Transactions on Graphics (TOG) 32 (6) (2013) 179.

[8] T. F. Cox, M. A. Cox, Multidimensional scaling, CRC press, 2000.

[9] J. Shawe-Taylor, C. K. Williams, N. Cristianini, J. Kandola, On the eigenspectrum of the gram matrix and the generalization error of kernel-pca, IEEE Transactions on Information Theory 51 (7) (2005) 2510–2522.

[10] M. E. Yumer, N. J. Mitra, Spectral style transfer for human motion between independent actions, ACM Transactions on Graphics (TOG) 35 (4) (2016) 137.

[11] F. Bach, R. Jenatton, J. Mairal, G. Obozinski, et al., Optimization with sparsity-inducing penalties, Foundations and Trends® in Machine Learning 4 (1) (2012) 1–106.

[12] S. J. Wright, R. D. Nowak, M. A. Figueiredo, Sparse reconstruction by separable approximation, IEEE Transactions on Signal Processing 57 (7) (2009) 2479–2493.

[13] J. Mairal, F. Bach, J. Ponce, G. Sapiro, Online dictionary learning for sparse coding, in: Proceedings of the 26th annual international conference on machine learning, ACM, 2009, pp. 689–696.

[14] D. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.

[15] E. Volkova, S. De La Rosa, H. H. Bülthoff, B. Mohler, The mpi emotional body expressions database for narrative scenarios, PloS one 9 (12) (2014) e113647.

[16] E. P. Volkova, B. J. Mohler, T. J. Dodds, J. Tesch, H. H. Bülthoff, Emotion categorization of body expressions in narrative scenarios, Frontiers in psychology 5.