TENSORIAL FACTORIZATION METHODS FOR MANIPULATION OF FACE VIDEOS

S. Manikandan, Ranjeeth Kumar, C.V. Jawahar

Center for Visual Information Technology
International Institute of Information Technology, Hyderabad
jawahar@iiit.ac.in

Keywords: Image Generation and Manipulation, Video Processing, Video Factorization, Face Morphing

Abstract

This paper proposes the use of Tensor Factorization for manipulating videos of human faces. Decomposition of a video represented as a tensor into non-negative rank-1 factors results in sparse and separable factors equivalent to a local parts decomposition of the object in the video. Such a decomposition can be used for tasks like expression transfer and face morphing. For instance, given a facial expression video it can be represented as a tensor which can then be factorized. The factors that best represent the expression can be identified which can then be transfered to another face video thus transferring the expression. A good solution to the problem of expression transfer would require explicit modeling of the expression and its interaction with the underlying face content. Instead the method proposed here is purely appearance based and the results demonstrate that the proposed method is a simple alternative to the popular complex solution. A similar strategy has been used to morph a face image in to a second face image. The resulting morph sequence was visually smooth indicating that the method can be used for generation of good quality morph sequence.

1 Introduction

Automated analysis of images or videos of human faces has been an important area of research in computer vision. The appearance of face is the most influential stimulus to perceptual systems that enables humans to identify and communicate with each other. The number of techniques developed for detection and recognition of faces in images demonstrates the strength of face as a biometric [14]. In addition to this, the ability to manipulate face videos to perform tasks like expression transfer, morphing has powerful applications in interactive systems, virtual worlds, gaming, video conferencing etc. The current work proposes methods to solve these problems using tensor factorization techniques. The problem of expression transfer is akin to the *translation* problem described in [8]. Given videos of subjects with certain expressions the goal of expression transfer is to synthesize videos of these subjects in all expressions present in the input. Figure 1 gives a visual definition of the expression transfer problem. There are number of applications of expression transfer: player-look-alike characters in computer games, personalized smileys that show person's image with expression rather than a generic animate face, interactive systems etc.

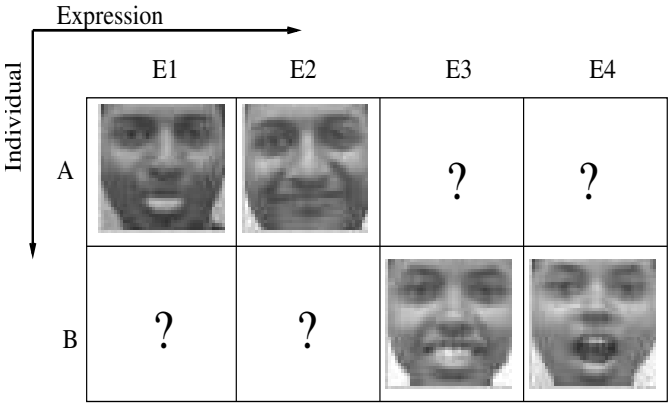

Figure 1: The expression transfer problem : Given some expression videos synthesize the missing ones (marked with ?).

A natural solution to the expression transfer problem would be similar to the style and content separation described in [8]. However modeling facial expression explicitly and its interaction with the face content is a very complex problem. Simple models such as the bilinear model [8] may not be rich enough to capture the interactions between these factors. Instead the method proposed here uses the local parts feature decomposition given by a positive factorization of a video represented as a tensor to carry out this task.

Another popular facial video manipulation task is morphing. The goal of face morphing problem is the generation of intermediate face images that depict a visually smooth transition from one face image to a second one. It had been widely used for generating visual effects. The method proposed in this work replaces the factors obtained by the 3D tensor factorization successively to obtain the intermediate images of the morph sequence. Section 2 describes earlier techniques on factorization, expression transfer and morphing. Section 3 describes tensorial factorization method that is used in this work. Sections 5 and 6 describe the algorithms and experiments followed by results and discussion in Section **??**.

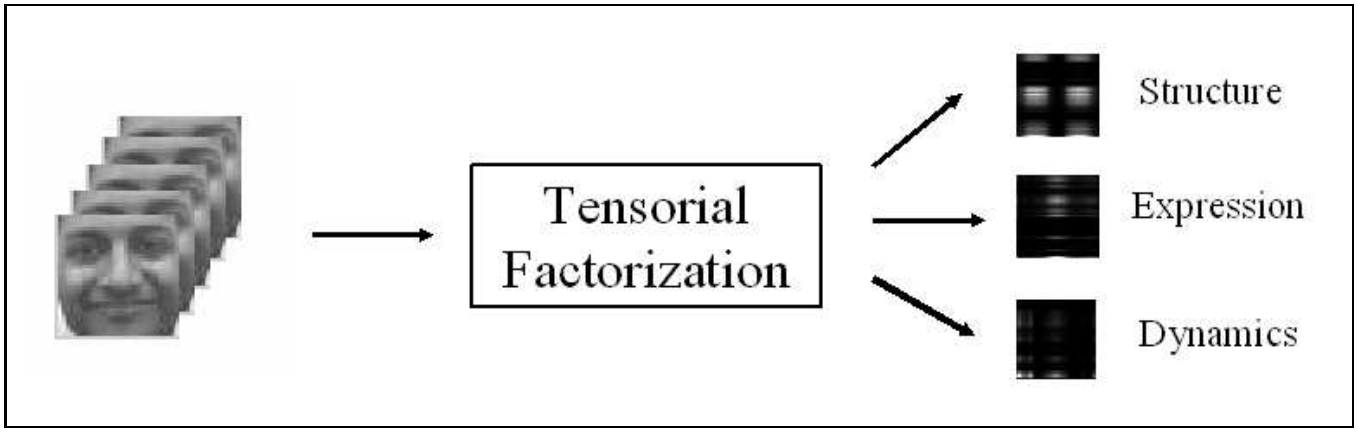

Figure 2: Tensorial Factorization of Videos: Video is treated as a tensor and decomposed into multiple factors. Factors represent various aspects of appearance and dynamics

2 Background

Factorization methods have been extensively used in computer vision for tasks like recovery of shape and motion [2], dimensionality reduction [12] etc. The methods use factorization methods like singular value decomposition (SVD) to decompose a measurement matrix into product of two matrices which represent properties like shape and motion. Decomposition methods similar to SVD have been proposed for tensors recently. A recent work [10] introduces the notion of tensorfaces analogous to the traditional eigenfaces based on multilinear analysis of image ensembles. They extend the traditional singular value decomposition (SVD) to an N-mode SVD. Such multilinear analysis helps us in extracting multiple factors that affect the appearance of an object in an image. Moreover, tensors are a natural representation of image ensembles or videos and such a representation retains the 2D structure of images thus preserving the spatial coherency in images. Recently alternative methods for tensor factorization have been proposed [1, 16] based on a positive preserving gradient descent scheme. The method results in factors that are sparse and separable unlike the factors obtained with the N-mode SVD. The central idea of the current work is to use the factors obtained by tensorial factorization of videos for the tasks of expression transfer and morphing.

Facial expression transfer has been studied extensively both in computer vision and graphics. The traditional warping based approaches like [9] ignore texture variations while morph based approaches [15] cannot be used to transfer an expression to a new face. A method based on ratio images was presented in [17] which can transfer expression as well as capture illumination variation. The problem of expression transfer requires not only separation factors like style (the expression) and content (the underlying face) but also a modeling of the interaction between such factors. Existing factors models like [5] have been found to be ineffective in capturing such interactions [8] . Bilinear models are used to model the interaction between style and content and model fitting is performed using a matrix factorization technique in [8]. The model has been successfully used for the problem of *translation* i.e translating new content in a new style to known content or known style. However they do not report results on transferring expressions. Du and Lin [4] attempt to learn a linear mapping between parameters representing expression and appearance. Wang and Ahuja [6] use Higher-Order SVD (*HOSVD*) as a multi-factor analysis method and decompose a collection of face expression images in to two separate expression and person subspaces and use them to map expression on to a new person's face.

The method proposed in this work does not attempt an explicit modeling of the interaction between the factors. Instead an input video is represented as a tensor and a set of sparse factors is extracted by decomposing the tensor. Factorizing a video in to positive rank-1 factors results has an interesting interpretation that the frames of the video can be considered as the linear combinations of a basis image set. Since the factors are all positive the basis images represent additive parts. Thus the basis set roughly corresponds to parts-based representation of objects in the image [3]. Figure 3 shows the basis obtained by decomposition of an image-cube. Observe that torso and the limbs of the *swimmer* get separated and code in different basis images [1]. Similarly, the decomposition of face results in parts like the nose, cheeks etc. Once such factors are extracted, the factors that best represent the expression are required for transfer of expression to another face video. However, in this work we do not address the problem of identification of such factors in detail. The expression-specific factors are presently identified using heuristic methods. These factors are then substituted in the video of the desired subject to synthesize a new video of with the expression transfered to another subject.

Earlier approaches to face morphing required the specification of feature points and the correspondence between them across

the two source and target images. A Bayesian framework for generating a morph field by distorting the brightness and geometry of the source image is proposed in [11]. The method proposed in the current work does not attempt to detect the feature points but replaces the factors obtained using the tensor factorization. The methods presented herein do not require feature point correspondences or complex modeling schemes. The simple appearance based approach is an efficient and robust alternative to the existing approaches, and generally achieves visually satisfactory results.

3 Tensorial Factorization

An N-valent tensor is an N-dimensional array. Given a video with frames of width w, height h and n frames, it can be represented as a $h \times w \times n$ tensor where every frame of the video forms a slice of the tensor. Let G be such a 3-valence tensor of dimensions $d_1 \times d_2 \times d_3$ indexed by the indices i_1, i_2, i_3 with $1 \leq i_j \leq d_j$. The rank of a higher order tensor G can be defined similar to the 2D (matrix) case. The tensor G is of rank at most k if it can be expressed as a sum of k rank-1 tensors, i.e a sum of 3-fold outer products:

$$G = \sum_{j=1}^{k} \mathbf{u}_1^j \otimes \mathbf{u}_2^j \otimes \mathbf{u}_3^j \tag{1}$$

where $\mathbf{u}_i^j \in R^{d_i}$. The rank of G is the smallest k for which such a decomposition exists. A decomposition of the tensor involves finding its rank-k approximation. While the notion of rank extends quite naturally to tensors finding the rank of a tensor or decomposing a tensor more difficult than in the matrix case. For matrices the rank-k approximation can be reduced to repeated rank-1 approximations while for tensors repeated deflation by dominant rank-1 tensors need not be a converging process. However the factorization of a tensor is usually unique unlike matrix factorization [7]. Algorithms such as the High-Order SVD (HOSVD) have been used for factorization of tensors. They are extensions of the traditional SVD method such that some of the features of SVD are preserved guaranteeing a reduction to SVD when tensor has the same image stacked repeatedly. A technique for factorizing tensors in to positive factors (PTF) by minimizing the reconstruction error is proposed in [13]. Unlike PCA or HOSVD algorithms, which result in factors that are not sparse, the PTF algorithm generates sparse factors owing to the positivity of factors [3]. Recently Hazan and Shashua [1] proposed a non-negative Tensor Factorization (NTF) method. Given a ND tensor G the method approximates G with a non-negative rank-k tensor $\sum_{j=1}^{k} \otimes_{i=1}^{N} \mathbf{u}_i^j$ described by Nk vectors $\mathbf{u}_i^j$ such that the reconstruction error:

$$\frac{1}{2} \left\| G - \sum_{j=1}^{k} \sum_{j=1}^{k} \otimes_{i=1}^{N} \mathbf{u}_i^j \right\|_F^2 \tag{2}$$

is minimized subject to $\mathbf{u}_i^j \geq 0$ where $\|A\|_F^2$ is the square Frobenius norm, i.e, the sum of squares of all entries of the tensor elements.

4 Factorization of Videos

The above described tensorial factorization can be applied to decompose videos. A 3D tensor is a very natural representation for videos. The 2D frames that compose the video are stacked to form a 3D tensor G. The method used is a gradient descent scheme with a mixture of Jacobi and Gauss-Seidel update scheme proposed in [1]. Taking the differentials of the reconstruction error with respect to the $\mathbf{u}_i^j$ and equating to 0 we obtain the following update rules for *i*-th component of the vectors $\mathbf{u}_1^j$:

$$u_{1,i}^j \leftarrow \frac{u_{1,i}^j \sum_{s,t} G_{i,s,t} u_{2,s}^j u_{3,t}^j}{\sum_{j=1}^{k} u_{1,i}^m \langle \mathbf{u}_2^m, \mathbf{u}_2^j \rangle \langle \mathbf{u}_3^m, \mathbf{u}_3^j \rangle} \tag{3}$$

and likewise for the vectors $\mathbf{u}_2^j$ and $\mathbf{u}_3^j$. It can be seen that these update rules preserve non-negativity provided the initial guesses for the vectors $\mathbf{u}_1^j$, $\mathbf{u}_2^j$ and $\mathbf{u}_3^j$ are positive. The update rules are shown be a converging procedure. Further details and a proof of convergence can be found [1] and [16]. Like PTF the method results in sparse and separable factors.

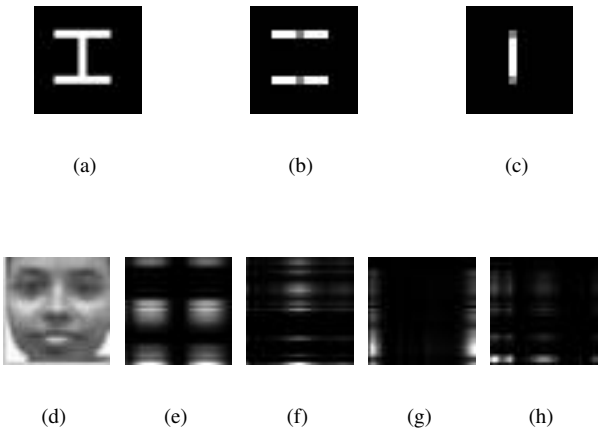

Figure 3: Result of tensorial factorization of a single image replicated 20 times to form a 3D tensor. The factors shown are the matrices $\mathbf{u}_1^i \otimes \mathbf{u}_2^j$ shown as images. Upper row: (a) the original synthetic image (b)-(c) the two recovered factors. Lower row: (d) the original image (e)-(h) the recovered factors in four groups

The factors resulting from the factorization of a video represent different aspects of the video such as the dynamics, appearance, structure etc. The outer product $\mathbf{u}_1^j \otimes \mathbf{u}_2^j \otimes \mathbf{u}_3^j$ can be interpreted differently when the matrices $\mathbf{u}_1^j \otimes \mathbf{u}_2^j$ are considered as images.

(a) (b)

Figure 4: (a) A frame of the original video (b) The corresponding frame in the video reconstructed using the factors obtained after decomposition

The coefficients from $\mathbf{u}_3^j$ combine these images to give the frames of the video. Thus these images describe the appearance while the vectors $\mathbf{u}_3^j$ encode the dynamics. Figure 3 shows the images obtained by factorizing image-cubes formed by stacking the same image and Figure 4 shows the reconstructed frame. Figure 5 shows the relative reconstruction error against the number of factors.

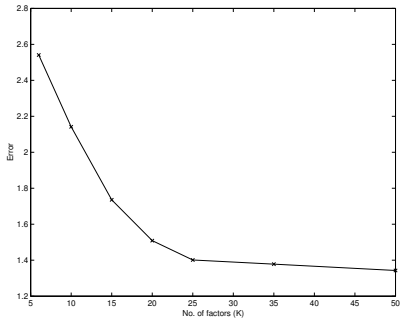

Figure 5: Graph showing plot of number of factors vs reconstruction error

5 Expression Transfer

As observed in [8] perceptual systems usually untangle the style and content factors from a single stimulus. The problem of expression transfer has a similar flavor where the expression can be thought of as style and the underlying face as content or vice-versa. Another way is to view the shape and texture characteristics of a face as the two factors contributing to the appearance of the face in the video. Separating those factors and modeling the interaction between them provides the ideal solution to this problem. The interaction between such factors may not be amenable to simple models like linear or bilinear models. As a simple and useful alternative we use the sparse factors obtained by a tensor factorization of the video and transfer the factors that best represent the expression to achieve expression transfer.

5.1 Algorithm

The expression transfer problem in its simplest form can be stated as: Given a video V_1 of person P_1 portraying an expression E_1 and a video V_2 of person P_2 portraying an expression E_2 synthesize videos of person P_1 portraying expression E_2 and of person P_2 portraying expression E_1. More generally given videos $V_1, \ldots, V_n$ of people $P_1, \ldots, P_m$ each depicting a corresponding set of expressions $E_1, \ldots, E_m$. Let E be the set of all expressions in the input i.e $E = \bigcup_{i=1}^{m} E_i$ and P be the set of all people in the input i.e $P = \bigcup_{i=1}^{m} P_i$. The system needs to learn all the expressions in E and then synthesize videos of person P_i portraying expressions in $E - E_i, \forall i = 1, \ldots, m$. Now given a video V_i of a person $P_i \in P$ depicting expression $e_j \in E$ create a tensor G_i using the frames of the video. Apply the described tensorial factorization scheme to G_i to decompose it into k factors $f_1, \ldots, f_k$. Analyze these factors to find out which of them represent the expression part of the video. Let F_{e_j} denote the subset of these factors that best represent the expression e_j. Obtain $F_{e_j}, \forall j = 1, \ldots, |E|$ in a similar manner. Now, to synthesize an expression of person P_l in an expression $e_j \in E - E_l$ using a video V_i (of that person) replace the corresponding expression factors of V_i with F_{e_j}.

The following are the main steps of the algorithm:

- Represent the input video as a tensor.
- Decompose tensor in to non-negative factors using the non-negative tensorial factorization method.
- Identify the factors that best represent the expression i.e. expression-specific factors.
- Use these factors to transfer expression by replacing expression-specific factors of one video with those of the second video.

The identification of expression-specific factors is an issue that needs further investigation. We do not address this intriguing issue in the current work. Instead, since the images $\mathbf{u}_1^j \otimes \mathbf{u}_2^j$ encode the appearance of the video we run over subsets of these images and choose those images that, when considered as filters, give widely varying responses over different frames of the video. We chose this heuristic since such images would represent the dynamic parts of the face whose appearance is specific to the expression. With a minimal manual intervention this method yields satisfactory results. A method for automatic identification of expression-specific factors using the neutral images of subjects is currently being developed.

We conducted experiments with a data set captured in house. The videos were taken using a Logitech webcam and consist of multiple subjects depicting different expressions. The videos were preprocessed such that the face occupied almost the entire

frame and the position of the face remained fixed. This was done by aligning the tip of the nose in every frame. The videos were then converted to grayscale and scaled to 40x40 pixels. Each video was of duration 4s @ 15fps for a total of 60 frames. Figure 6 shows the results of the proposed algorithm for expression transfer.

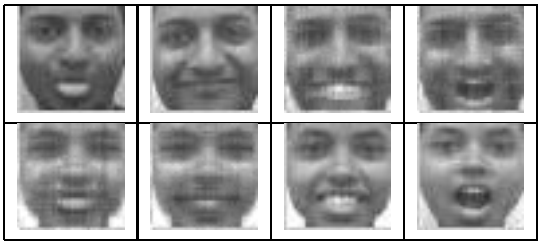

Figure 6: Transfer of expressions. The matrix shown in Figure 1 completed using the proposed algorithm. The synthesized videos along with the original input videos. Note that the transfer was done using appearance clues alone without the use of any feature correspondences or muscular motion models.

6 Application to Face Morphing

Image metamorphosis also known as morphing consists of a fluid transformation from a source image to target image. The technique has several applications in generating visual effects and recognition of faces. Traditional techniques for generation of morph sequences require image warping and color interpolation. 2D geometric transformation are applied to retain the alignment between the images while color interpolation blends the colors. Once again, this can be viewed as a smooth transition in the shape and texture characteristics of the image. Factorizing the video provides a natural way to generate such a morph sequence by replacing the factors of source video with the factors of the target video successively.

We show how tensorial factorization can be used to perform facial morphing. Given images of two faces, the system needs to synthesize images of the transition from one face to another. To achieve this, we represent the first image as a 3D tensor by stacking the image multiple times. We now perform tensorial factorization on this tensor and retrieve the factors that represent this face. We repeat this process for the second face image to retrieve the factors that represent the second image. We now generate the transition faces by replacing one factor of the first image by the corresponding factor from the second image and reconstructing the tensor using the new set of factors. Any frame of this reconstructed tensor can be used as the intermediate image (as all the frames will be the same). This gives the first transition image. The above process can be repeated successively transferring all the factors to generate all the required transition images. We will get as many transition images as the number of factors with this method.

The following steps constitute the morphing algorithm:

- Represent the first face image as a tensor by stacking the image multiple times. Build a second tensor similarly using the second face image.
- Factorize both the tensors to get the factors representing each of them.
- Replace the first factor of the first tensor with the corresponding factor from the second tensor.
- Reconstruct the first transition image using the new factors.

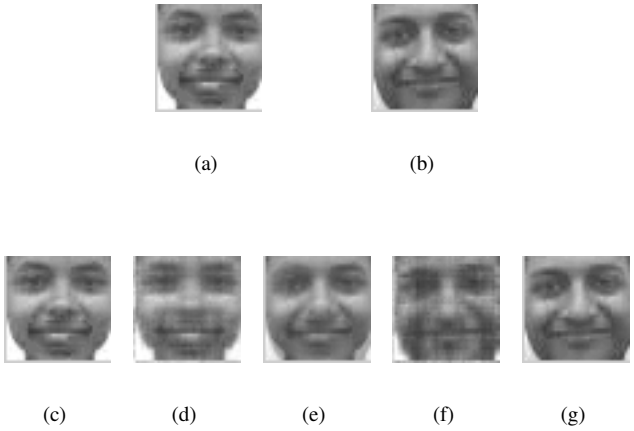

Figure 7: Face Morphing. Upper row: (a), (b) the two input images. Lower row: (c), (g) the input images, (d) - (f) the transition images.

- Repeat the above process to obtain the other transition images.

Figure 6 shows the result of applying this algorithm on a pair of faces. It can be seen that the transition is visually smooth. Thus the algorithm provides a simple yet effective way to perform face morphing. Apart from the applications described above video factorization has promising prospects in activity recognition. Tensorial factorization can be viewed as separation of the appearance and kinematics of an object performing an activity. Efforts to use this representation for tasks like activity recognition/detection are in progress.

7 Discussion

The results of performing expression transfer using the algorithm described in in section 5 are shown in figure 6. Figures 1 and 6 shows the screen shots of the videos that were input to the system and screen shots of the videos that were

synthesized by the system. The ?s in figure 1 indicate those videos that have to be synthesized by the system. Screen shots of these videos are shown in the lower table. While the results demonstrate that it is possible to transfer expressions in this manner it must be noted that the transfer is not achieved by separation of expression and facial content factors. Instead the appearance of the expression in one video had been transfered to the other and consequently the appearance of the synthesized video would largely depend on the source video as different subjects articulate the same expression differently. However, it is clear from figure 6 the synthesized videos are visually satisfactory.

The results of the morphing using the algorithm described in section 6are shown in the figure 6. Figure 7(a) and 7(b) were given as the two input images to the system. The transition images generated by the system are shown in figures 7(d) - 7(f). The successive replacement of factors results in smooth transition from source face image to the target image. It must be noted replacing the factors directly is a naive way of obtaining the transition. Possible improvements include achieving smoother transition images by ranking factors from two videos based on similarity and the transferring the factors so as to reduce discontinuities in the transition.

8 Conclusion

We have used used positive factorization of videos represented as tensors for the problems of expression transfer and morphing demonstrating that such factorization can be used for facial video manipulation effectively. Performance is demonstrated on a set of in house captured videos. The decomposition of videos into multiple factors, appears to be a promising direction for analyzing and understanding dynamic events in videos. Identification of factors specific to a video, using the representation for analysis of activities in videos are promising directions for future work.

References

[1] Amnon Shashua and Tamir Hazan. Non-negative tensor factorization with applications to statistics and computer vision. *Proc. of the International Conf. on Machine Learning*, 2005.

[2] Carlo Tomasi and Takeo Kanade. Shape and motion without depth. In *Proc. of the Third IEEE International Conf. on Computer Vision*, pages 91–95, 1990.

[3] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[4] Du Y. and Lin X. Mapping emotional status to facial expressions. In *Proc. of the International Conf. on Pattern Recognition*, volume II, pages 524–527, 2002.

[5] Geoffrey E. Hinton and Zoubin Ghahramani. Generative models for discovering sparse distributed representations. *Philosophical Transactions Royal Society*, 352(1177–1190), 1997.

[6] Hongcheng Wang and Narendra Ahuja. Facial expression decomposition. In *Proc. of the Ninth IEEE International Conf. on Computer Vision*, page 958, 2003.

[7] J. Kruskal. Three way arrays: Rank and uniqueness of trinlinear decomposition with applications to arithmetic complexity and statistics. *Linear Algebra and its Applications*, 18:95–138, 1977.

[8] Joshua B. Tenenbaum and William T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 12:1247–1283, 2000.

[9] Jun-yong Noh and Ulrich Neumann. Expression cloning. In *Proc. of the Annual Conf. on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 277–288, 2001.

[10] M. Alex, O. Vasilescu, and Demetri Terzopoulos. Multilinear analysis of image ensembles: Tensorfaces. In *Proc. of the Seventh European Conf. on Computer Vision*, volume 1, pages 447–460, 2002.

[11] Martin Bichsel. Automatic interpolation and recognition of face images by morphing. *Proc. of the Second International Conf. on Automatic Face and Gesture Recognition*, pages 128–135, 1996.

[12] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of Cognitive Neurosccience*, 3(1):71–86, 1991.

[13] Max Welling and Markus Weber. Positive tensor factorization. *Pattern Recognition Letters*, 22:1255–1261, 2001.

[14] Stan Z. Li and Anil K. Jain, editors. *Handbook of Face Recognition*. Springer, New York, USA, 2005.

[15] Steven M. Seitz and Charles R. Dyer. View morphing. In *Proc. of the Annual Conf. on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 21–30, 1996.

[16] Tamir Hazan, Simon Polak, and Amnon Shashua. Sparse image coding using a 3d non-negative tensor factorization. *Proc. of the Tenth IEEE International Conf. on Computer Vision*, 2005.

[17] Zicheng Liu, Ying Shan, and Zhengyou Zhang. Expressive expression mapping with ratio images. In *Proc. of the 28th Annual Conf. on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 271–276, 2001.