Realistic Face Animation for Audiovisual Speech Applications: A Densification Approach Driven by Sparse Stereo Meshes

Marie-Odile Berger, Jonathan Ponroy, Brigitte Wrobel-Dautcourt

Realistic face animation for audiovisual speech applications: a densification approach driven by sparse stereo meshes

Marie-Odile Berger, Jonathan Ponroy, and Brigitte Wrobel

LORIA/INRIA Nancy Grand Est
Marie-Odile.berger@loria.fr,Brigitte.Wrobel@loria.fr

Abstract. Being able to produce realistic facial animation is crucial for many speech applications in language learning technologies. Reaching realism needs to acquire and to animate dense 3D models of the face which are often acquired with 3D scanners. However, acquiring the dynamics of the speech from 3D scans is difficult as the acquisition time generally allows only sustained sounds to be recorded. On the contrary, acquiring the speech dynamics on a sparse set of points is easy using a stereovision recording a talker with markers painted on his/her face. In this paper, we propose an approach to animate a very realistic dense talking head which makes use of a reduced set of 3D dense meshes acquired for sustained sounds as well as the speech dynamics learned on a talker painted with white markers. The contributions of the paper are twofold: We first propose an appropriate principal component analysis (PCA) with missing data techniques in order to compute the basic modes of the speech dynamics despite possible unobservable points in the sparse meshes obtained by the stereovision system. We then propose a method for densifying the modes, that is a method for computing the dense modes for spatial animation from the sparse modes learned by the stereovision system. Examples prove the effectiveness of the approach and the high realism obtained with our method.

Key words: Face animation, densification, PCA with missing data

1 Introduction

There is a strong evidence that the view of speaker's face visual information noticeably improves the speech intelligibility. Hence, having a realistic talking head could help language learning technology in giving the student a feedback on how to change articulation in order to achieve a correct pronunciation. In [7], Munhall and Vatikiotis provide evidence that lip and jaw motions affect the entire facial structure below the eyes. High levels of details are thus required to obtain highly realistic and perceptibly correct facial animation of the complete face.

Though the utility of visual information to speech perception has been known for a long time, progresses to develop talking faces which both look real and

convey linguistic relevant information are more slower. What causes the intelligibility enhancement afforded by the visible component of speech is difficult to determine. For these reasons, many works argued the necessity for animating face directly from visible articulatory data either in 2D [3] or in 3D [2, 4–6]. Most of these data are 3D meshes of the face which can be sparse -when they are acquired using markers or sensors glued on the face- or dense using laser scanner acquisition. It must be noted that do not operate with a sufficient speed for speech acquisition (an acquisition rate of 120 Hz is required to acquire fast articulatory gestures of consonants). As a result, high resolution scanners can be used for sustained sounds as vowels but it is not obvious that they can be used to acquire all the dynamics of speech production.

Most methods for animating face are based on the extraction of the basic modes of spatial deformations during speech using principal component analysis (PCA). They span a space which describes at best all the plausible face deformations corresponding to speech production. Constructing such a space requires to physically match the points of the meshes at each time instant. This task is easy when sparse meshes are considered but it becomes complex for dense meshes. In this latter case, a generic head template is generally used [5, 6] to align all the scans using prominent features such eyes, nose. Obtaining rough alignment is largely automatic but accurate alignment requires a manual adaptation of the model to the speaker specificities. In this paper, we favor a fully automatic method and propose an automatic matching process guided by the sparse stereo meshes.

Though PCA can be used both on sparse or dense meshes, it is difficult to obtain the speech dynamics from dense meshes since the set of visemes that can be acquired is limited to sustained sounds due to technical limitations in scanner acquisition. It is thus not obvious to recover the complete speech dynamics from a reduced set of 3D scans. For these reasons, we propose in this paper an approach that makes use both of a small set of 3D scans acquired for sustained sounds and of the speech dynamics learned from a stereovision sequence of the face painted with markers. The main idea is to transfer the dynamics learned on the sparse meshes onto the 3D dense meshes in order to generate realistic dense animations of the face. This paper is an extension of our previous work [1]. In this past work, only one dense mesh was used which turned out to be not accurate enough for generating sounds which are too far from the reference dense mesh. We thus propose in this paper significant extensions of the work in order to transfer the dynamics learned from the sparse meshes onto the dense face.

Computing PCA modes for face animation from sparse meshes acquired by stereovision is difficult because all the markers may not be observable at every time instant. When the mouth is closed, points may become invisible, or at least not sufficiently visible, to be correctly detected and reconstructed. Though this difficulty is often ignored in the literature, this make the PCA more complex and needs to resort to PCA with missing data techniques [8]. A novel way of computing the face modes taking into account possible unobservable points is thus presented in section 3.

Our approach for computing the dense modes is presented in section 4. It borrows concepts from transfer techniques [10] used in computer graphics to map an object onto another. We specifically used this technique in section to physically and automatically match the dense meshes using the underlying sparse mesh to guide the transfer. This avoids doing manual adaptation of a generic mesh to our talker. Finally, we propose a densification method which allows the basic modes of the dense head to be computed from the sparse modes computed from the learning sequence. As a result, this allows us to animate a dense head using only a reduced set of 3D scans.

We demonstrate experimentally in section 5 that the proposed method is efficient and allows us to obtain very realistic dense faces.

2 System overview

Our method requires the acquisition of a set of 3D dense meshes of the talker. In our study, these dense meshes were acquired with the Inspeck mega capturor (www.inspeck.com) for 15 sustained sounds (vowels, fricatives and lip closure). These dense meshes are also called visemes in the following.

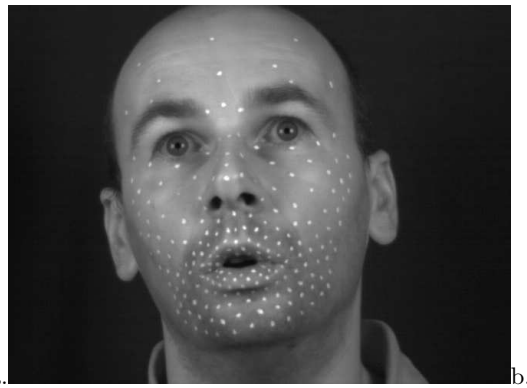
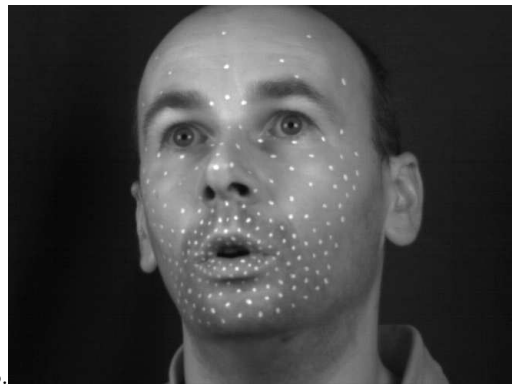
In order to learn the face kinematics, a classical stereovision system with two cameras was used to record a corpus. The acquisition rate of the cameras is 120 images/frames which is sufficient to capture fast movements of the articulators (further details on this system can be found in [11]). Markers were painted on the talker's face in order to make the matching and the reconstruction stage automatic. With 45 points on the lips and a total of 209 points on the part of the face that is influenced by speech, the recovery of face kinematics is quite detailed. Our experiments proves that between 5 to 7 PCA modes are sufficient to describe the face kinematics.

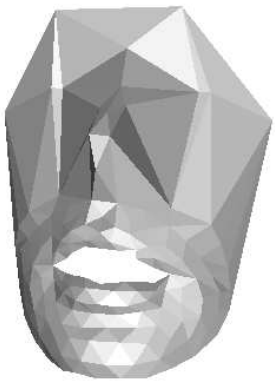
The experimental set-up and the input data are shown in Fig.1. The first raw exhibits a stereo image pair of the talker. Note that the points located on the top of the head are used to compensate for head motions. Fig. 1.c is an example of a sparse mesh obtained with the stereo system. Finally, Fig.1.d is the dense mesh acquired for the /a/ sound.

3 PCA with missing data for computing the sparse modes

The 3D coordinates of each marker can theoretically be computed at each time instant of the learning sequence using the stereovision system. In the following, the 3D sparse mesh of the face computed at time instant t ($t \in [1..T]$) are denoted $X_t = [X_{1,t}, ..., X_{N,t}]$, where N is the number of markers and $X_{i,t}$ are the 3D coordinates of the marker i at time instant t. Here, N=209 markers were painted on the face. The duration of the corpus recorded for learning kinematics was 6 minutes, giving rise to T=39000 stereo pairs.

However, some markers may become unobservable during uttering when the lips are very close. Practically, these markers may not be reliably detected in the

a.  b. 



c.

d.

Fig. 1. Input data of the system: (a and b) a couple of stereovision images. White points/markers were painted on the face ; (c): reconstruction of the sparse mesh; (d): the 3D dense map obtained for the /a/ sound.

images by low level algorithms. It is especially the case for sounds like /u/, /o/, /i/. Markers may also become unobservable in one of the image pair due to slight head motions which make some points disappear from the field of view. Stereo reconstruction is thus not possible for these points. Practically speaking, 77% of the markers are always observable and thus reconstructed. 17% of the markers have a reconstruction rate in the range $[70\%, 95\%]$ and 6% of the markers have a reconstruction rate less than 50% and are mainly located on lips (Fig .2). As a result, some data are may be missing in the X_t data.

Most of the algorithms for building principal subspaces are based on the decomposition of the covariance matrix of the input data and cannot be used when some data are missing, as in the present case. However, there exist probabilistic approaches where PCA is considered as a limiting case of a linear Gaussian model. Principal axis can then be computed using Expectation-Maximisation (EM) algorithms [8]. Such algorithms can be extended to handle the problem of missing data [9, 8].

In this paper, we adapt these ideas to our particular problem in order to compute the sparse modes. These ideas are also used in section 4 to compute the dense modes. We first compute the principal components for the markers $\mathcal{R}$ which are always reconstructed over the sequence. The components can be easily computed with classical methods since all the data are available for these markers. We then complete the entries of the principal components by introducing the markers which are not observable at every time instant using EM techniques.

Let $\{u_k^r\}_{k \leq q} \in \mathbb{R}^r$, be the q principal components computed from the set of markers which are always observable and let $\bar{X}^r$ be the mean of these meshes. Let $u \in \mathbb{R}^N$ be the extended basis we are looking for. Given a mesh X_t acquired at time instant t, let X_t^r be the reduced mesh, where only the markers always reconstructed are considered. X_t^r can be approximated on the q principal components as:

$$X_t^r \approx \bar{X}^r + \sum_{k=1}^{q} \alpha_{k,t} u_k^r \quad t = 1..T \tag{1}$$

The goal of the complete components is to approximate any mesh as a linear combination:

$$X_t \approx \bar{X} + \sum_{k=1}^{q} \beta_{k,t} u_k \quad t = 1..T \tag{2}$$

As X_t^r is a sub-vector of X_t - 77% of entries of X_t are the entries of X_t^r in our set-up- , it is likely that $\beta_k = \alpha_k$ is very close to the mean square solution of (2). In the same way, we can consider that the entries of $\bar{X}$ which correspond to always reconstructed markers are identical to the entries of $\bar{X}^r$.

We then build a linear system which incorporates all the observations on the non always observable markers. For each observation of a marker i at time instant t, equation (2) gives rise to the following linear equation :

$$X_{i,t} = \bar{X}_i + \sum_{k=1}^{q} \alpha_{k,t} u_{k,i}$$

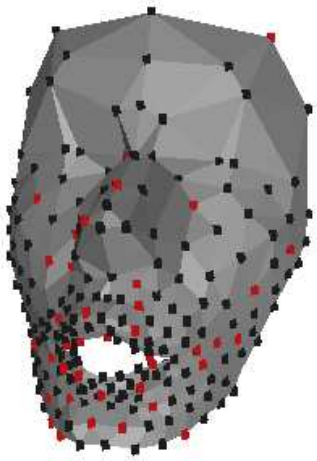


Fig. 2. The markers in black are the ones which are reconstructed in every frame and are taken into account in the reduced PCA.The markers in red are not always reconstructed in the sequence and are considered in the computation of the principal components using an iterative EM algorithm.

Stacking all the equations for each visible marker at each time instant t leads to a linear system where the unknowns are the missing $u_{k,i}$ ($i \notin \mathcal{R}$) and the missing components of the mean $\bar{X}$. The $\alpha_{k,t}$ are computed by projecting the reduced meshes onto the reduced components $\{u_k^r\}$. Considering all the markers which can be reconstructed at some instant of the sequence, we obtained a set of linear equations which can be solved in the least square sense, giving rise to the missing entries of the principal components and of the mean.

The principal components are then refined using the classical EM algorithm:

- Given the components u_k and $\bar{X}$, compute the coefficients $\alpha^{t,i}$ for all the meshes available in the sequence.
- given the $\alpha_{t,i}$, refine the $u_{k,i}$ and $\bar{X}$ by solving for each i the system of linear equations:

$$X_{i,t} = \bar{X}_i + \sum_{k=1}^{q} \alpha_{k,t} u_{k,i}, \quad t = 1..T$$

in the least square sense.

4 Generating dense modes from sparse modes

Dynamic realism is needed for linguistic expressions. Achieving realism needs to have a high resolution 3D model as well as knowledge about the facial kinematics. In our case, high resolution scanners of the face can only be obtained for sustained sounds as vowels. We thus propose in this paper to transfer the change in shape exhibited on the sparse meshes onto the dense mesh and to infer dense animation modes from the sparse ones.

Here gain, as in the computation of the sparse mode, we use a densification process to compute the dense modes from the sparse ones.

4.1 Overview

Let $X_1^{dense}, .., X_d^{dense}$ be the set of d dense visemes which were scanned. This set contains 15 visemes of the french vowels as well as a neutral expression and a face with closed mouth.

We first suppose that these meshes physically matched, which means that the vertex i in every mesh fits the same physical point in all the dense meshes. Additionally, we also suppose that the dense and the sparse meshes are registered. How to obtain such data will be considered in the next section.

The aim of PCA is to express each viseme as a linear combination of the dense modes:

$$X_i^{dense} \approx \bar{X}^{dense} + \sum_{k=1..q} \beta_k^i u_k^{dense}, i = 1..d \tag{3}$$

The space spanned by the sparse modes is built so as to describe the plausible appearances of a face. There thus exists a sparse mesh X_i which best fits each dense mesh X_i^{dense}. Hence, we can compute coefficient $\alpha_{k,i}$ such that:

$$X_i \approx \bar{X} + \sum_{k=1}^{q} \alpha_{k,i} u_k \tag{4}$$

where u_k is the set of sparse modes computed in section 3.

Unlike the preceding case, X is not dense in X^{dense}: the number of vertices of the sparse mesh is 209 whereas the dense mesh contains around 13000 vertices. However the vertices of the sparse mesh are distributed on the whole face and vertices are more present in mobile areas of the face during articulation. For these reasons, we also consider that $\beta_k = \alpha_k$ can be considered as a fair least square solution of equation (3).

These first estimates of β are then use to solve (3) in the least mean sense in the unknown modes u_k^{dense} and the mean $\bar{X}^{dense}$. Note that we estimate less than 7 modes. As we have 15 dense visemes, the system always has a solution.

4.2 Obtaining physically matched dense visemes

Dense visemes are recorded with a laser scanner and the recovered meshes do not physically match. Remeshing of the visemes must thus be performed.

However, physically matching points between deformable surfaces is known as a very difficult problem. We thus take advantage of the underlying sparse model to obtain a physically coherent parametrization.

The first step is to identify the sparse mesh which best fits each dense viseme. As the sparse shapes are described using a reduced number of sparse modes, this can be done by computing the rigid displacement which minimizes the distance between the dense mesh and the sparse mesh thank to an iterative closest point algorithm. We are thus looking for the displacement $\mathcal{T}$ and the coefficient α which minimize:

$$Min_{\mathcal{T},\alpha_1,..,\alpha_q} distance(\mathcal{T}^{-1}(\bar{X} + \sum \alpha_k u_k), Dense\ Mesh)$$

Doing this, we both identify the sparse mesh which best fits the dense viseme as well as the displacement between the two surfaces. For sake of simplicity, the registered dense visemes $\mathcal{T}(V^{dense})$ are still denoted V^{dense} in the following.

It is important to note that the registered visemes contains all the vertices of the sparse model. This property is used to physically match the dense visemes under the control of the sparse meshes. We here consider the /a/ viseme as the reference mesh. Remeshing is achieved with respect to this reference viseme and is performed on the basis of an extended affine matching between each pair of corresponding facets for the sparse visemes.

Given a facet ABC of the sparse reference viseme and its 3D position $A'B'C'$ in the viseme to be matched (see figure 5), let M be a point of the reference dense visem. Each point of the dense mesh is associated to a sparse facet using the algorithm we described in [1]. The main idea is to define M as a function of the vertices of the facet and of the normals. The line HM, with H belonging to the facet, is affinely defined with respect to the facet: H is defined as the point with affine coordinates $(\alpha, \beta, 1-\alpha-\beta)$ such that HM and $\alpha N_a + \beta N_b + (1-\alpha-\beta)N_c$ are collinear. We then define $H'M'$ as the line with the same affine coordinates with respect to the facet $A'B'C'$ and its normals. M' is finally obtained as the intersection of this line with the dense mesh.

5 Results

Figure 4 shows the first dense mode computed from the sparse mode. Fig 4.a shows the mean shape of the sparse meshes and Fig. 4.b the computed dense mean shape. The second row exhibits the sparse first modes as well as the computed corresponding dense modes.

Figure 5 shows examples of dense faces computed from the corresponding sparse meshes. Given faces randomly taken within the sparse meshes in the database, the coefficients of the sparse mesh onto the sparse mode were computed. The dense face was then reconstructed from these coefficients using the

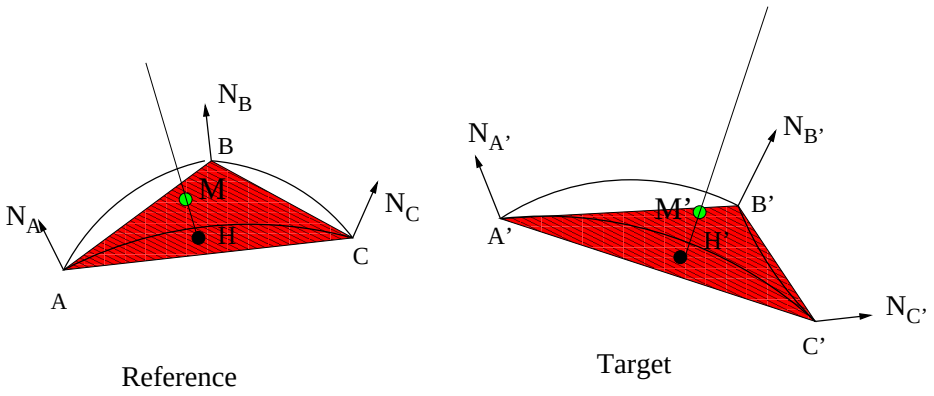


Fig. 3. How a physically registered dense mesh is obtained using the underlying sparse meshes.

computed dense modes. The first row 5 shows the sparse mesh and the corresponding dense face. Other reconstructed dense faces are shown in the second and third row. These examples prove the high realism of the obtained face with a global motion of the face under the eyes coherent with the speech gesture. A full dense sequence is available on the webpage http://www.loria.fr/~berger/teteParlante. Given a speech sequence of our corpus, the PCA coefficients were computed using the sparse modes and the dense head was generated for each image. The visual impression of the resulting sequence is very good and proved that our method is able to produce realistic facial animation.

It must be noted that the reconstruction is somehow incomplete in the inner region of the lips. This is due to the fact that the transfer procedure only allows the dense face in the neighborhood of the sparse mesh to be computed. We plan to investigate in future works methods to complete the lips from the acquired dense meshes. Lip modeling as realized in [5] is a possible research direction.

6 Conclusion

We have proposed an approach that produces automatically highly realistic face animation. This method does not require expensive materials: two cameras are required for the stereovision system and only a reduced set of 3D scans of the talker are needed. The main strength of the approach is to transfer the speech dynamics which can be easily learned on sparse data onto the dense face. To this aim, we have proposed an original densification approach which allows to compute the dense modes from the sparse modes. This method produces highly realistic dense postures. In the near future, we plan to conduct perceptive evaluation of the head with the aim to prove that this very realistic face improves speech intelligibility.

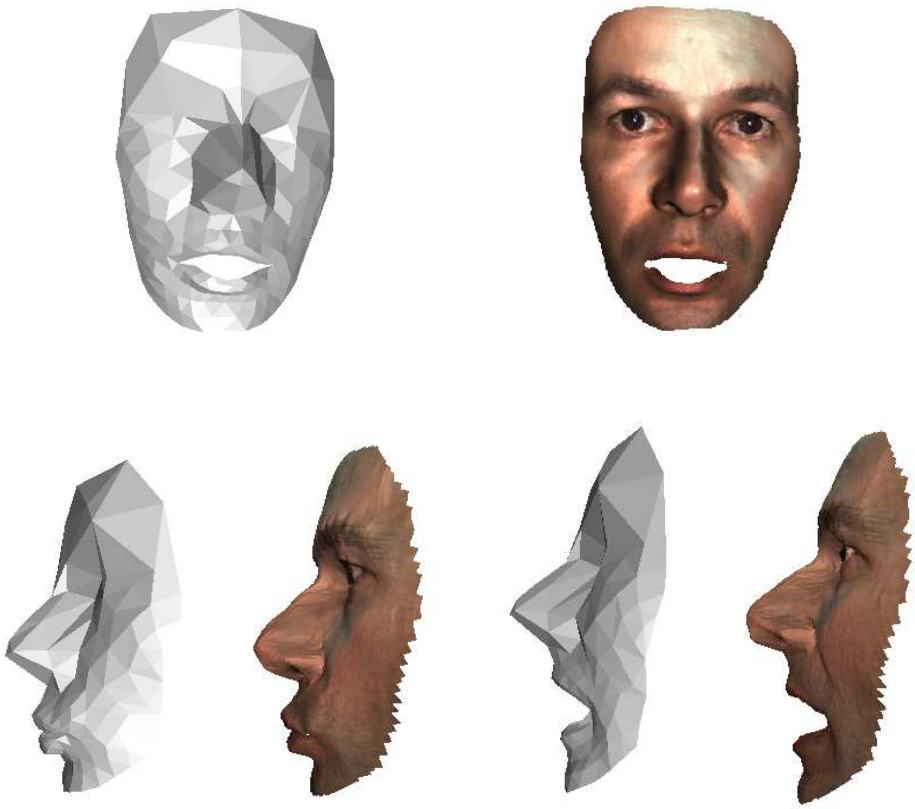


Fig. 4. First row: the mean shape of the sparse meshes and the computed dense mean mode. Second row: $mean - 3 \times u_1$ and $mean + 3 \times u_1$ are exhibited for the sparse modes and the computed dense modes.

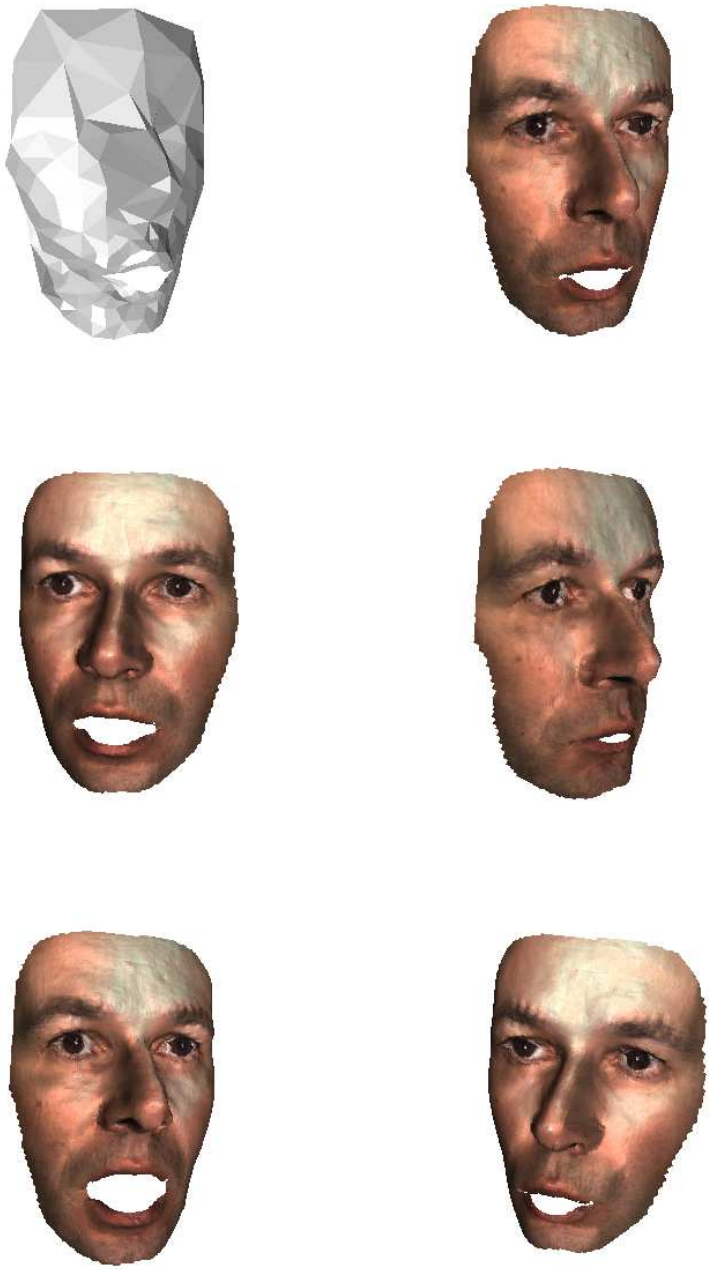


Fig. 5. Examples of dense faces generated using 7 dense modes.First row: a sparse mesh and the corresponding dense mesh. Second and third row: examples of dense faces computed from various sparse meshes.

References

1. M.O. Berger. Realistic face animation from sparse stereo meshes. In *International Conference on Auditory-Visual Speech Processing 2007*, 2007.
2. V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH 99*, pages 187–194, 1999.
3. Tony Ezzat, Gadi Geiger, and Tomaso Poggio. Trainable videorealistic speech animation. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 388–398. ACM Press, 2002.
4. G. Kalberer and L. Van Gool. Face animation based on observed 3d speech dynamics. In *Proceedings of Computer Animation 2001 Conference*, pages 20–27, 2001.
5. Takaaki Kuratate. *Talking Head Animation System Driven by Facial Motion Mapping and a 3D face Database*. PhD thesis, Nara Institute of Science and Technology, 2004.
6. P. Muller, G. Kalberer, M. Proesmans, and L. Van Gool. Realistic speech animation based on observed 3d face dynamics. *Vision Image and Signal Processing*, 152(4), 2005.
7. K. Munhall and E. Vatikiotis-Bateson. The moving face during speech communication. In *Hearing by Eyes, volume2, chapter6, Psychology press*, pages 123–139, 1998.
8. S. Roweis. Em algorithms for pca and spca. In *Advances in Neural Information Processing Systems*, pages 626–632. IEEE and ACM, 1998.
9. Danijel Skocaj, Horst Bischof, and Ales Leonardis. A robust pca algorithm for building representations from panoramic images. In *ECCV (4)*, pages 761–775, 2002.
10. R. Summer and J. Popovic. Deformable transfer for triangle meshes. In *Procceding of SIGGRAPH 04*, 2004.
11. B. Wrobel-Dautcourt, M.-O. Berger, B. Potard, Y. Laprie, and S. Ouni. A low-cost stereovision based system for acquisition of visible articulatory data. In *Proceedings of the 5th Conference on Auditory-Visual Speech Processing (AVSP), Vancouver Island, BC, Canada, July*, pages 65–70, 2005.