Realistic Face Animation From Sparse Stereo Meshes

Marie-Odile Berger[1]

[1]INRIA-Lorraine/LORIA, 615 rue du jardin Botanique, BP 101, 54602 Villers -les-Nancy, France

berger@loria.fr

Abstract

Being able to produce realistic facial animation is crucial for many speech applications in language learning technologies. For reaching realism, it is necessary to acquire and to animate dense 3D models of the face. Recovering dense models is often achieved using stereovision techniques. Unfortunately, reconstruction artifacts are common and are mainly due to the difficulty to match points on untextured areas of the face between images. In this paper, we propose a robust and fully automatic method to produce realistic dense animation. Our input data are a dense 3D mesh of the talker obtained for one viseme as well as a corpus of stereo sequences of a talker painted with markers that allows the face kinematics to be learned. The main contribution of the paper is to transfer the kinematics learned on a sparse mesh onto the 3D dense mesh, thus allowing dense facial animation. Examples of face animations are provided which prove the reliability of the proposed method.

Index Terms: modeling of facial gesture, face animation.

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. This task is complex and necessitates a multidisciplinary effort involving speech production modeling and image analysis. Within the scope of improving the learning capacities of a foreign language, we are here interested in building realistic animations of a tutor given a text as input. We are thus especially interested in recovering realistic dense modes of face deformations. In [1], 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 realistic facial animation.

Modeling facial dynamics is essential for creating realistic animations but it is difficult to achieve due to limitations in current shape capture technologies. Very few shape capture methods work effectively for rapidly moving scenes. In particular laser scanner techniques exist but they do not operate with a sufficient speed for speech acquisition (an acquisition rate of 120Hz is required to acquire fast articulatory gestures in consonants). Among shape capture methods, only stereovision techniques are able to acquire 3D dense sets with the required acquisition speed [2, 3, 4].

Given two or more calibrated camera and given two corresponding points in the images (i.e points that correspond to the same physical point), the 3D point is built as the intersection of the two rays that pass though the optical centers and the image points. The key problem to be solved in stereovision is the matching stage where points that correspond to the same physical 3D point must be identified in the two images based on similar intensity or color. This problem is highly difficult and can be simplified by projecting light patterns on the face [3] in order to acquire a dense map. Unfortunately, such a technique does not capture motion, since points do not physically correspond over time.

In order to make the matching stage easier, markers can be glued or painted on the face. Such markers can be easily detected and matched in the images. This method only allows a sparse map of the face to be obtained and is thus of limited interest for face reconstruction/modeling/synthesis. However, this method is widely used to build a dynamic model of the face using principal component analysis [5, 4, 6, 7].

Obtaining a dense map of the face from stereovision techniques is much more difficult, especially because identifying corresponding points between stereo images is difficult except for particular points (eyes, eyebrows,...). As a result, classical reconstruction methods are not robust and false matchings are often present, especially at depth discontinuities and in regions presenting near uniform texture. To cope with these problems, stereo reconstruction is now often considered as the minimization of a cost functional that integrates spatial and temporal regularization constraints [2, 8]. As for most regularization methods, the parameters of the regularization functions need to be carefully tuned to obtain satisfying results in order to avoid reconstruction artifacts or excessive smoothing. In addition, only a set of unstructured points is obtained: a mesh must then be inferred from the set of 3D points in order to render and to animate the model. Finally, these methods are very computational demanding: in [2] the computation is distributed over multiple machines to speed up the process.

In order to reduce interaction and parameter tuning in the acquisition process and to improve the robustness of the process, we propose an approach that borrows concepts from model-based approach and from vision based tracking of markers on the face. Our input data are a dense 3D map of the talker obtained for **one** viseme as well a a corpus of stereo sequences of the talker with markers painted on his/her face that allow the face kinematics to be computed. In this study, the 3D dense map was acquired with a scanner for a sustained vowel but other acquisition technologies can be used. The main idea of the paper is to transfer kinematics learned on the sparse mesh onto the 3D dense mesh in order to generate realistic dense animations of the face. As a result, we are able to recover dynamic dense meshes that do not present depth artifacts nor over-smoothing. As a side effect, we are also able to recover the dense modes of the talker, thus allowing easy animation of the head.

Our experimental set-up is presented is section 2. The transfer method is presented in section 3. Finally experimental results are exhibited and discussed in section 4.

2. The experimental set-up

Our method requires the acquisition of one 3D dense mesh of the talker. In our study, this dense mesh was acquired with the Inspeck mega capturor (www.inspeck.com) for the sustained /a/ sound because the lips are fully visible for this sound. In order to learn the face kinematics, a classical stereovision system with two cameras has been used to record a corpus. The acquisition rate of the cameras is 120 images/frames which is sufficient to capture fast movements of the articlulators (further details on this system can be found in [7]). Markers were painted on the face of the talker in order to make automatic the matching and the reconstruction stage. With 45 points on the lips and a total of 200 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. Fig.1.a is a snapshot of the acquisition set up for the sparse corpus. Fig.1.b and c are two stereo images of the talker. Note that the points on the top of the head are used to compensate for head motions. Fig. 1.d is an example of the sparse mesh obtained with the stereo system. Finally, Fig.1.e is the dense mesh acquired for the /a/ sound.

In order to express the sparse and the dense meshes in the same frame coordinate, the two meshes are registered using the iterative closest point algorithm (ICP). Specifically, as the vertices of the sparse mesh belong to the face, registration is performed by computed the displacement R,T which minimizes:

$$Min_{R,T} \sum_{M \in Sparse\ Mesh} d(RM + T, Dense\ Mesh)$$

3. Transferring sparse face dynamics onto the 3D dense mesh

Given a 3D dense mesh and the face kinematics, our goal is now to transfer the change in shape exhibited on the sparse meshes onto the dense mesh during speech articulation. The problem of deformation transfer plays a central role in computer graphics and aims at applying the deformation exhibited by a source onto a different target object [9, 10]. Such techniques are commonly used in animation to reproduce body animation from one character to another. Our particular application is in some sense a transfer problem: the source is here the sparse mesh kinematics and the problem is to reproduce on the 3D dense mesh the motion defined by the sparse mesh (Fig. 2).

Deformation transfer was used for expression cloning in [9] for animation purpose. In this approach, a mapping between the vertices of the source and the target was required and a radial based function was used to approximate the mapping between the two surfaces globally. Heuristics were then used to adapt the direction and scale of displacement vectors to account for local variations of shape and proportion between faces. This approach is well suited to animation of characters because only a rough visual agreement is required. Unfortunately, such global modeling of face deformations is not suited to our study because high fidelity to local gestures and constraints of speech articulation is needed.

Representing the deformation as a collection of affine transformations defined on each facet of the mesh was advocated in[10] because deformations can be locally controlled and global non linear deformations of the surface can be easily considered. It is important to note that correspondences of the three vertices of a facet before and after deformation do not fully determine the affine transformation since they do not establish how the space perpendicular to the facet deforms. In [10], a fourth undeformed vertex perpendicular to the facet and at a distance 1 was defined to allow computation of the full affine transformation. As a consequence, the transformation in the space perpendicular to the facet preserves distances which restricts the bending properties of the surface.

In the present work, we start from the local affine parameterization and extend it to our particular purpose by considering an affine framework both for the matching step and the deformation of the facets. Each point of the dense mesh is matched with one of the facet of the reference sparse mesh. This point is then affinely defined with respect to the facet. The coordinates of this physical point for a new sound are then recovered as the point with the same affine coordinates with respect to the corresponding facet of the target sound. Our main contributions are two-fold:

- we define an automatic affine mapping process between the dense mesh and the sparse mesh,
- we compute a complete affine deformation for each facet, especially in the space perpendicular to the facet.

3.1. Affine mapping of the sparse and the dense mesh

Matching curves or surfaces - i.e find point to point correspondences between two structures- is still considered as a difficult problem. The main difficulty is to find a matching process that preserves the topological and differential properties of the surfaces. To only mention some desirable properties, the process must be continuous (two nearby points on the surface are matched with nearby points of the other), the pairing must align local shapes (loosely speaking, this means that flat regions are matched with flat regions, strongly curved regions are matched with strongly curved regions) and smooth regions must be matched with smooth unfolded regions .

The aim of this step is to affinely map the dense mesh onto the sparse mesh in order to later transfer the sparse motion onto the dense mesh. For sake of clarity, the mapping process is first explained in 2D (Fig. 3). The discrete normal at each vertex A of the sparse mesh is first computed and is denoted as N_A in the following. The 2D space around the curve can thus be partitioned in cells. Each point around the sparse mesh can thus be associated with a unique cell except if the curve is almost self intersecting or presents rapid curvature variations. In this latter case, the point is affected to the nearest facet. Thinks can be easily extended to 3D space using 3D cells.

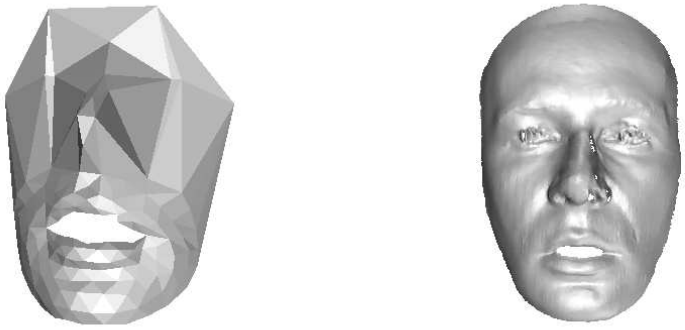
Given a point M on the dense mesh, its corresponding point H on the sparse mesh is computed with an affine projection. H is defined as the point with affine coordinates $(\alpha, \beta, 1-\alpha-\beta)$ such that HM and $\alpha N_a + \beta N_b + (1-\alpha-\beta)N_c$ are collinear. This way, a regular matching is induced between the dense and the sparse mesh that respect the common points between the two meshes. To get a better idea of this matching stage, the points of the dense mesh associated to the same triangular facet of the sparse map are drawn with the same color in Fig. 3 (below). This figure proves that the matching process is reliable.

3.2. Defining the transfer process

Let us recall that the transfer function is represented as a collection of affine transformations defined on each facet of the mesh.

Given a facet ABC of the reference sparse mesh and its position $A'B'C'$ in the target sparse mesh, we want to compute

a. b. c.



d. e.

Figure 1: Input data of the system: (a):The stereovision system; (b and c) a couple of stereovision images ; (d): reconstruction of the sparse mesh; (e): the 3D dense map obtained for the /a/ sound.

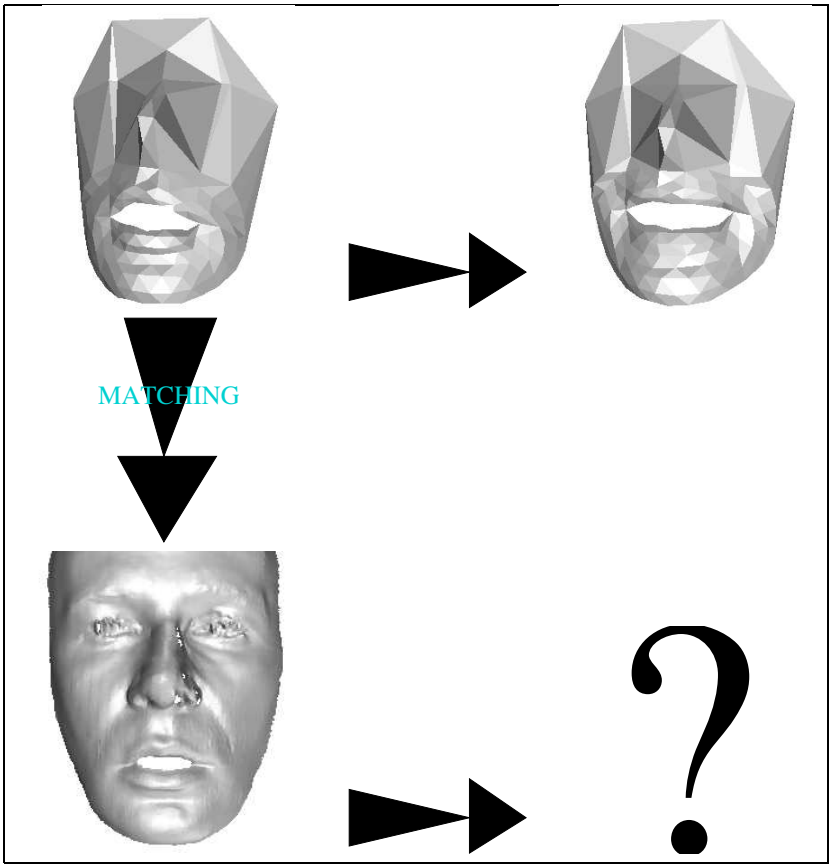


Figure 2: The transfer problem: reproduce the kinematics of the sparse mesh on the dense mesh.

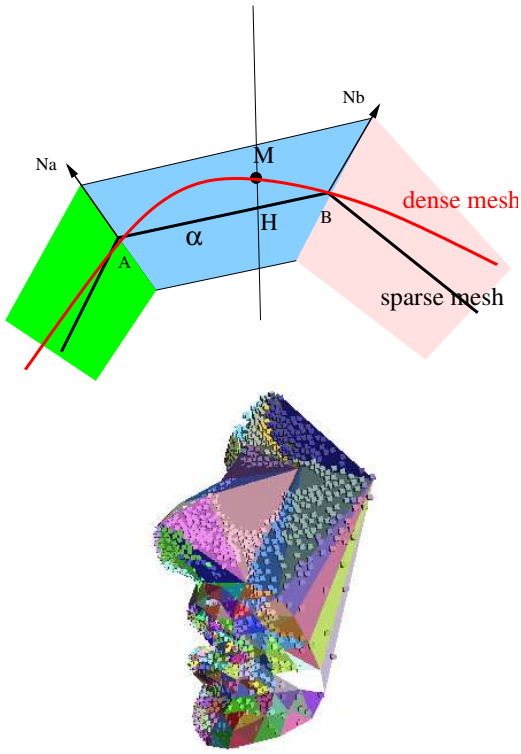


Figure 3: Matching the dense with the sparse mesh. a:Affine projection in 2D, b: the matching stage.

for each point M associated with ABC its position $f(M)$ after motion. f is known at the vertices of the facets because the points of the sparse mesh correspond over time:

$$f(A) = A',\ f(B) = B'\ and\ f(C) = C' \qquad (1)$$

f is thus defined for each point of the facet ABC. Defining the complete transfer function needs to define what appends in the space perpendicular to ABC (Fig. 4). As the kinematics is defined by the evolution of a triangular mesh, it seems natural to define the transfer function by taking into account not only the correspondences between points but also the evolution of the curvature of the surface between the target and the reference sparse mesh. For a mesh surface, the curvature at a vertex P is defined as [11]:

$$K = 2\pi - \sum \alpha_i$$

where α_i are the angles between successive edges at P. This means that the triangular surface can be approximated by a sphere with radii K. K encodes information for each vertex on the bending of the surface. Imposing that the curvature vectors correspond yields additional constraints on f:

$$\begin{array}{l} f(K_A N_A) = K'_A N'_A \\ f(K_B N_B) = K'_B N_B \\ f(K_C N_C) = K'_C N_C \end{array} \qquad (2)$$

These equations convey information not only on the considered facet but also on the neighboring facets.

Beeing affine, f has 12 degrees of freedom. After having imposed constraints on vertices correspondences, there are still 3 degrees of freedom. Using one curvature constraint is theoretically sufficient to solve for f. However, in order to take into account all the curvature information we compute f as the affine transformation that exactly satisfies equation(1) and that satisfies the curvature constraints (2) in the least square sense. The solution to this linear constrained optimization problem can be obtained using classical singular value decomposition techniques (SVD) [12].

Each 3D point of the dense map associated with the facet ABC is then transfered using the computed affine function f.

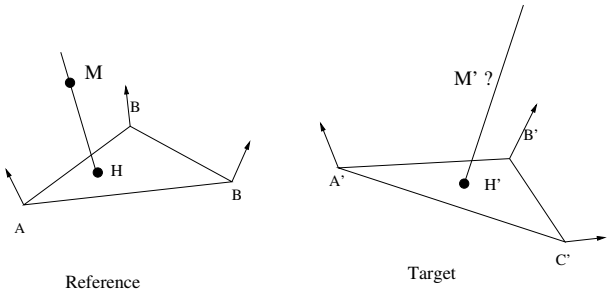


Figure 4: The transfer process.

4. Results

As a result, given one dense mesh of the talker for a reference sound, any dense map for another target sound considered in the sparse sequences can be recovered. In our experiments the /a/ sound was chosen as the reference. The sparse mesh and the dense mesh are exhibited in Fig. 5.a and Fig. 5.d. The dense mesh transfer was then tested on the /aka/ speech sequence. Some sparse meshes are shown (Fig.5.b and c) and the computed dense meshes are exhibited in (Fig.5.e and f).

Using the same reference sound, results of mesh transfer are shown in Fig. 6 on the /any/ speech sequence.

The visual impression of the recovered dense meshes is very good and the method is able to produce realistic expressions of the whole face.

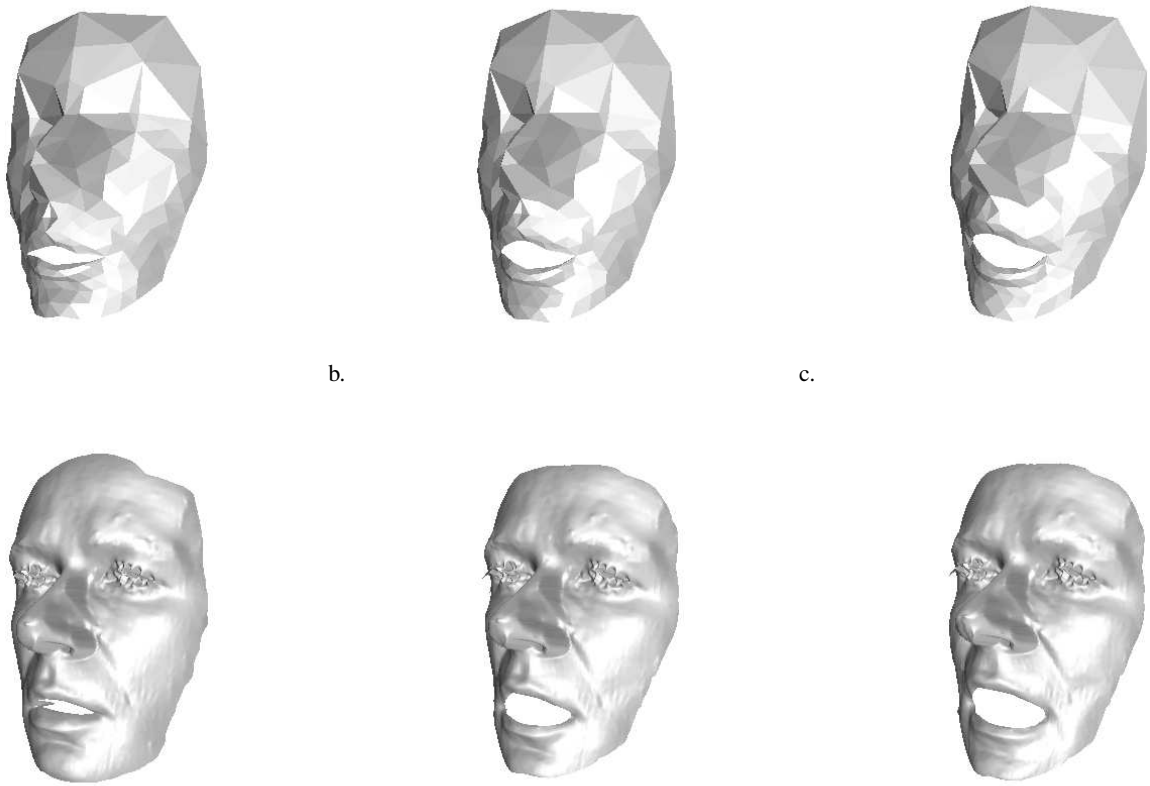
As the lips are not fully visible on the sparse map, transfer is unable to give a realistic animation of the complete lips. That is the reason why results only exhibit the parts of the lips that were visible on the sparse data. We plan to complete the dense mesh with a lip model in the near future.

An important application of our method concerns the computation of the modes of the dense mesh. Indeed, the transfer process can be used to transfer the modes of the PCA computed from the sparse data onto the dense data. This allows the dense modes to be obtained, thus enabling the animation of the dense talking head.

5. Conclusion

A parameter free robust method for realistic dense mesh animation has been presented in this paper. This method does not require expensive materials: two cameras are needed for kinematics acquisition and only one dense mesh of the talker is needed. Its primary interest is to produce realistic animations of the whole face without requiring any parameter tuning. Our first results are very promising.

Perceptive evaluation of the process will be conducted in the very near future. Our aim is to determine to what extent this realistic face animation improves speech intelligibility.



a. b. c.

d. e. f.

Figure 5: Examples of dense face generation during the /aka/ sound: (a) is the reference sparse mesh and (d) is the corresponding dense model. (d) and (f) are the dense meshes generated from the sparse meshes b and c.

Figure 6: Dense face generation during the /any/ sound.

6. References

[1] K. Munhall and E. Vatikiotis-Bateson, "The moving face during speech communication," in *Hearing by Eyes, volume2, chapter6, Psychology press*, 1998, pp. 123–139.

[2] L. Zhang, N. Snavely, B. Curless, and S. M. Seitz, "Spacetime faces: High-resolution capture for modeling and animation," in *ACM Annual Conference on Computer Graphics*, August 2004, pp. 548–558.

[3] P. Huang, C. Zhang, and F.-P. Chiang, "High-speed 3-d shape measurement based on digital fringe projection," *Optical Engineering*, vol. 42, no. 1, pp. 163–168, 2003.

[4] G. Kalberer and L. V. Gool, "Face animation based on observed 3d speech dynamics," in *Proceedings of Computer Animation 2001 Conference*, 2001, pp. 20–27.

[5] F. Elisei, M. Odisio, G. Bailly, and P. Badin, "Creating and controlling video-realistic talking heads," in *Proceedings of Audio-Visual Speech Processing, Aalborg, Denmark, 2001.*, 2001.

[6] V. Robert, B. Wrobel-Dautcourt, Y. Laprie, and A. Bonneau, "Strategies of labial coarticulation," in *Proceedings of the 9th European Conference on Speech Communication and Technology Interspeech'2005 - Eurospeech, Lisboa, September*, 2005.

[7] 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*, 2005.

[8] J.-P. Pons, R. Keriven, O. Faugeras, and G. Hermosillo, "Variational stereovision and 3d scene flow estimation with statistical similarity measures," in *International Conference on Computer Vision*, 2003, pp. 597– 602.

[9] J. Noh and U. Neumann, "Expression cloning," in *Proceedings of ACM SIGGRAPH 2001*, 2001, pp. 403–410.

[10] R. Summer and J. Popovic, "Deformable transfer for triangle meshes," in *Procceding of SIGGRAPH 04*, 2004.

[11] M. Meyer, M. Desbrun, P. Schroder, and A. Barr, "Discrete differential geometry operators for triangulated 2-manifolds," in *2002. VisMath.*, 2002.

[12] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C, The Art of Scientific Computing*. Cambridge University Press, 1988.