AN HYBRID IMAGE PROCESSING APPROACH TO LIPTRACKING INDEPENDENT OF HEAD ORIENTATION

L. Revéret[1], F. Garcia[2], C. Benoît[1], E. Vatikiotis-Bateson[2]
Institut de la Communication Parlée
(1) INPG/ENSERG/Université Stendhal, BP25 38040 Cedex 9 Grenoble, France.
Tel. +33 4 76 82 41 28, FAX: +33 4 76 82 43 35
(2) HIP-ATR Laboratories, 2-2 Hikaridai, Seika-sho, Soraku-gun, Kyoto, 619-02, Japan
Tel. +81 774 95 1011, FAX: +81 774 95 1008
E-mail: {reveret, benoit}@icp.grenet.fr, {garcia, bateson}@hip.atr.co.jp

ABSTRACT

This paper examines the influence of head orientation in liptracking. There are two main conclusions: First, lip gesture analysis and head movement correction should be processed independently. Second, the measurement of articulatory parameters may be corrupted by head movement if it is performed directly at the pixel level. We thus propose an innovative technique of liptracking which relies on a "3D active contour" model of the lips controlled by articulatory parameters. The 3D model is projected onto the image of a speaking face through a camera model, thus allowing spatial re-orientation of the head. Liptracking is then performed by automatic adjustment of the control parameters, independently of head orientation. The final objective of our study is to apply a pixel-based method to detect head orientation. Nevertheless, we consider that head motion and lip gestures are detected by different processes, whether cognitive (by humans) or computational (by machines). Due to this, we decided to first develop and evaluate orientation-free liptracking through a non video-based head motion detection technique which is here presented.

1. INTRODUCTION

Being able to see a speaker's face can dramatically increase the intelligibility of an auditorily degraded message. Of the many speech articulators that may be used in "lip-reading" (e.g., teeth, tongue, chin, etc.), the lips are clearly crucial to both the production and the perception of visible speech. The speaker's head is continuously moving under natural conditions of speech communication. Head movements are associated with pragmatic and prosodic information. Head motion always occurs, even under highly constrained condition, since the head cannot be physically fixed. Human perceivers seem able to extract phonetic information from lip gestures despite changes in orientation caused by head motion. Furthermore, it appears that sufficient information to decode lip gestures can be retrieved without foveating directly on the speaker's lips, suggesting that the required visual information is distributed dynamically at relatively low spatial frequency [11]. Although our understanding of this remarkable process of human perception is still very poor, we may easily assume that lip-reading requires two relatively independent cognitive processes: dynamic detection of the lip contour is likely to be an essential component and one whose implementation in automatic speechreading systems can be used to extract machine parameters most relevant to the phonetic decoding of visual speech.

In this paper, we present an original approach to liptracking combining techniques developed at ATR in Japan and at the ICP in France. Lip contours are parametrized by a 3D model developed at the ICP [3], which is actively fitted to the inner and outer contours of the speaker's lips. Although previously tested only on head-stabilized lip motion, an inherent feature of this 3D model is that the orientation angles (e.g., rotations cutting the face plane) can be used to predict (within reasonable limits) the speaker's lip geometry, regardless of the relative position of the head in the camera field of view. For obvious practical reasons, the final objective of our study is to detect head motion reliably using image-based techniques. However, a more complex technique is used at first in order to evaluate our liptracking approach without introducing head motion biases in our data. So that head orientation could be accurately measured from an independent source, we recorded the 3D position of several OPTOTRAK ired markers attached to the speaker's face during video-recording. The three rotation angles (pitch, yaw, roll) were derived from the xyz positions of the markers while the subject nodded and then wagged his head. The active contour could then be projected onto the speaker's face, whatever its orientation angle.

2. INFLUENCE OF HEAD MOTION IN LIP-READING SYSTEMS

2.1. Approaches to automatic lip-reading

Current lip-reading systems can be divided into two main classes of model-based and image-based systems. In the model-based approach, a geometrical contour is applied to the mouth area and adjusted to fit the inner and/or outer (active) contours of the lips. Splines [4] or

polynomial equations [12] are commonly used. The ASM (*Active Shape Model*) method adds a statistical constraint to the model to regularize its shape variation [2, 5]. Image-based systems directly process images at the pixel level, with no *a priori* knowledge of the lip shapes. These methods include both statistical approaches such as texture segmentation [6, 9] and dynamical approaches such as Optical Flow Analysis (OFA) [8]. Although several of the methods listed above allow correction for limited head movements, major head motion often corrupts the measurement process.

2.2. Head motion characteristics

The overall motion of the head with regard to a fixed system (typically the camera system) can be described by 6 degrees of freedom (figure 1.): Three translations (x, y and z) and the three rotations around these axes. We describe the inherent limitations of the various lip-reading systems and the corrections that need to be applied.

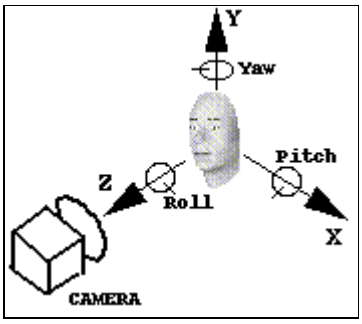

Figure 1. The six degrees of freedom of the head

2.2.1. Model-based approach

"Active contour" models can be adjusted to some head motion contained in the camera plane (roll, x, y) through a geometric transformation [4]. "Active Shape" models may also be adapted to some roll and translation [2, 5]. A scaling correction is also performed to account for the perspective effect due to movement along the z axis. However, these methods do not handle the problem that changes in perspective are highly similar to those due to lip protrusion. Another drawback in ASM techniques is that, since statistics on gray scale images mostly reflect changes in the vertical position of the lip contours [5], they are poorly sensitive to horizontal head motion.

2.2.2. Image-based approach

In these approaches, yaw, roll and z translation induce changes in lip orientation and scaling that may distort measurement. Without *a priori* knowledge of the lip geometry, no easy correction can be applied. First, confusion between scaling and protrusion remains. Second, in approaches based on texture segmentation, height and width of the contours might be underestimated if a roll movement occurs. Head motion is a critical problem in dynamic image-based approaches (such as OFA) where speech-related lip movements and head gestures are mixed in the same flow, making it impossible to separate the two sources of motion.

2.2.3. Problems common to the two approaches

With the two above mentioned methods, articulatory reliable interpretation of results is difficult: Distortions of lip shapes due to involuntary head motion introduce strong errors in the measurements of characteristic parameters (e.g., contour height or width). In addition, when taken into account, head motion is often estimated from the actual lip motion, although both "deformations" are superimposed. We here advocate that the head and lip movements should be processed independently. This is why we below present a brief overview of head motion detection, independent of internal deformations (such as expressions or lip gestures). Finally, we consider that the facial images should be seen as an intermediate observation space, rather than as a direct measurement space.

3. HEAD MOTION DETECTION

3.1. Current approaches

The importance of head motion measurement has increased over the last decades due to the development of human-machine interfaces entailing face detection or recognition, and identification of expressions under unconstrained conditions [10]. Detection of facial features (elliptic shape, skin color, face templates, etc.) often serves as a basis for detecting head location in an image.

[1] proposes a method based on OFA to extract parametrization of the rigid motion in terms of the translations and rotations. From a previous position of the head, this method estimates the current position by minimizing the difference between the velocity field measured from one image to the other and the velocity field generated by an ellipsoidal model of the head from the estimated actual position to the next one. So far, our estimates of head motion from OFA of video sequences of speaking and moving heads have not been accurate enough for reliable decomposition of head tracking and liptracking. Hence, our decision to use another means to measure head motion independently of lip motion.

3.2. Optotrak system

In order to extract accurate data that can serve as the reference against which other techniques are judged, we used the Optotrak system to measure head motion. This system detects in real time the absolute position of infra-red markers. The output is a list of xyz values in the coordinate space of the Optotrak camera. In our experiment, five markers were attached to the head of the speaker through a rigid wooden structure. First, the

center of rotation of the movement is defined. Then, a rigid body decomposition can be deduced from the recording. The rigid motion of the body defined by the markers is finally described in terms of three rotations and three translations. The center of rotation of the head was automatically identified by a regression technique. The regression algorithm converges towards a point which maintains constant distances between each marker and this point. Such a constraint is characteristic of a rigid motion around one articulation point.

4. Our approach to orientation-free liptracking

We propose a solution to liptracking, based on a novel approach to active contour modeling, that obeys the two requirements suggested in § 2. Head motion and lip gestures are processed independently using measurements not deduced directly from the screen representation. Instead of using 2D contours of the lips for liptracking, we use the projection of a 3D lip model on the camera plane.

4.1. The ICP lip model

A 3D model of the lips has been developed at the ICP to model the mouth gestures related to expressionless production of French speech [3]. At the highest command level, this model is controlled by five geometrical parameters : the inner contour height and with, protrusion of the inner contact point, and the protrusion of the upper and lower lips. These control parameters do not result from physical commands, nor do they correspond to the degrees of freedom of lip gestures. However, they are easy to measure [6], and even to predict [7]. Moreover, intelligibility tests of the model have shown speech recognition by humans is enhanced in noisy conditions.

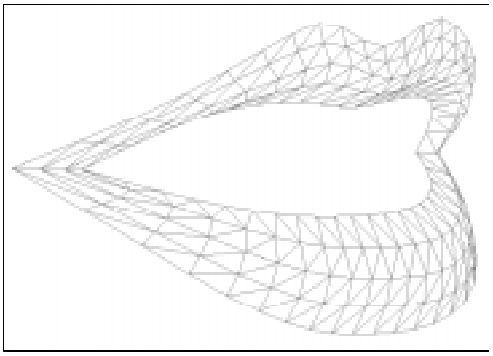

Figure 2. The 3D lip model developed at the ICP

4.2. 3D Active contour

In traditional methods based on active contour, distortion of the model is controlled by internal parameters of the analytic description: polynomial coefficients [12], control points for the spline model [4]. Then, for a given image of the lips, these control parameters are calculated by an optimization algorithm that maximizes the gradient along the pixels of the contour. Some regularization term may be introduced to smooth the deformations of the model. Once an extremum is found, the articulatory parameters are deduced from the resulting shape of the model. In the ICP model, the articulatory parameters themselves are the control parameters.

4.2.1. Orientation of the 3D lip model

Let $L0 = \{L0_i \in \Re^3\}$ be the set of 3D points that defines the lip model for the zero head position when the subject is looking directly into the camera. This initial position has no rotations and no translations. L0 depends on the five lip parameters P. Head position is described by a position vector Ω including the three rotations and the three translations. From Ω are calculated a rotation matrix R and a translation vector T. From any position Ω, the position of the model L_i is then given by

$$L_i(\Omega, P) = R(\Omega)\, L0_i(P) + T(\Omega)$$

4.2.2. Projection through a camera model

The 3D model is projected onto the screen through a camera model. To take into account perspective artifacts, a thin-lens camera model has been introduced. A 3D point (xyz) originally defined in the camera-centered system of coordinates is projected through this camera model (C) onto a 2D point (uv) = C(xyz) as:

$$u = x / (1 + z / f)$$
$$v = y / (1 + z / f)$$

where f is the focal length of the camera. Since the camera could be misaligned with the subject, the position $\Omega_c = (R_c, T_c)$ defines the position of the camera in the Optotrak coordinate system. The position of the camera being known, the projection Y on the screen of any point X defined in the Optotrak coordinate system is given by

$$Y(u,v) = C\,[\, R_c^{-1}(X - T_c)\,] = S(X)$$

Figure 3 shows the application of the camera model on the 3D lip model and on a simplified model of the subject's head.

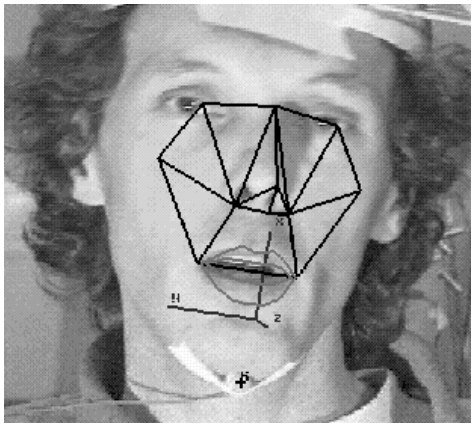

Figure 3. Projection of the lip and head models

A calibration procedure sets the internal parameters f and Ω_c of the camera. This procedure minimizes the differences between the position of a set of markers whose 3D Optotrak coordinates are known, and their projection calculated by the camera model.

4.2.3. The tracking algorithm

As for the 2D active contour, the position of the model is set to stick as much as possible to the area of highest gradient. Only the lip parameters Pi control the distortion of the model. The 2D contour is the projection of the lip model through the camera model, after rotation and translation. The convergence algorithm is ultimately formulated as

$$P^* = \text{argmax} \left(\Sigma_i G[S(X_i(P))] \right)$$

where G is the spatial gradient at the pixel position.

5. CONCLUSION

We have presented a liptracking method totally independent of head orientation, which allows liptracking from various viewing angles including the profile. Currently, an independent head tracking system is required to estimate accurate positions of the head. In the future, this invasive method will be replaced by purely image-based processing techniques.

REFERENCES

[1] Basu, S., Essa, I., Pentland, A., "Motion regularization for model-based head tracking", Technical Report 362, MIT Media Laboratory, Perceptual Computing Section, Jan. 1996.

[2] Garcia, F., Vatikiotis-Bateson, E., "Active Shape Model for Lip Tracking", in Proc. *ATR Symposium on Face and Object Recognition*, Jan. 20-23, Kyoto, Japan, pp. 58-59, 1997.

[3] Guiard-Marigny, T., Adjoudani, A., Benoit, C., "A 3D model of the lips for speech synthesis", *in Progress in Speech Synthesis*, J. Van Santen (eds.), Springer-Verlag, 1996.

[4] Kaucic, R. , Dalton, B. , Blake, "Real-Time Lip Tracking for Audio-Visual Speech Recognition Applications", in Proc. *ECCV*, pp. 376-387, Cambridge, UK, 1996.

[5] Luettin, J., Thacker, N. A., Beet, S. W., "Speechreading using Shape and Intensity Information", in Proc. *4th ICSLP Conference*, Philadelphia, PA, USA, 1996.

[6] Lallouache, M.T., "Un poste visage-parole couleur. Acquisition et traitement automatique des contours des lèvres", PhD. dissertation, INPG, Grenoble, France, 1991.

[7] Le Goff, B., "Automatic modeling of coarticulation in text-to-audiovisual speech synthesis", Proc. *the EUROSPEECH Conference*, Rhodes, Greece, this volume, 1997.

[8] Mase, K. , Pentland, A., "Automatic Lipreading by Optical-Flow Analysis", *in Systems and Computers in Japan*, vol. 22, no. 66, pp. 67-75, 1991.

[9] Petajan, E. , Graf, H. P., "Robust Face Feature Analysis for Automatic Speachreading and Character Animation", in *Speechreading by Man and Machine*, D. Stork and M. Hennecke Eds., Springer-Verlag, Berlin, pp. 425-436, 1996.

[10] Shapiro, L. S., Brady M., Zisserman, A., "Tracking Moving Heads", in *Real-Time Computer Vision*, Newton Institute, Cambridge University Press, 1994.

[11] Vatikiotis-Bateson, E., Munhall, K.G., Hirayama, M., Kasahara, Y., Yehia, H., "Physiology-Based Synthesis Of Audiovisual Speech", Proc. *the ESCA Workshop on Speech Production Modeling*, May 20-24, Autrans, France, pp. 241-244, 1996.

[12] Yuille, A.L., Hallinan, P.W., Cohen, D.S., "Feature Extraction from Faces using Deformable Templates", *Int. J Computer Vision*, vol. 8, no. 2, pp. 99-111, 1992.