POSE COMPENSATION FOR BIMODAL SPEECH RECOGNITION

Jianbo Ma

ECE Dept, University of Illinois
Urbana, IL 61801
mjb@vision.ai.uiuc.edu

Chalapathy Neti and Andrew W. Senior

IBM T.J.Watson Research Center
P.O.Box 218, Yorktown Heights, NY 10598
{cneti, aws}@ibm.us.com

ABSTRACT

Lip reading has been proven to improve speech recognition accuracy in adverse environments. Most existing lip reading systems have frontal pose assumption, which makes it very difficult to use in tasks such as video transcription (speech recognition of the audio stream for video indexing and retrieval). In this paper, we propose a new method to compensate the lip pose change by exploiting the general symmetry of human face. From the imaging geometry we show that a frontal lip can be recovered from only one profile view. The resulting pose compensation method has the following advantages: (1) it only requires one profile image; (2) it does not need any 3D model; (3) it does not need an accurate lip shape contour. Experimental results are given to show the effectiveness of our method.

1. INTRODUCTION

A good example of utilizing audio-visual interaction for human speech communication is lip reading [3]. Lip reading has been proven to improve speech recognition accuracy in adverse environments [9, 6]. In the lip reading process, visual features (lip image features) are detected, then these features are either combined with audio features to train an audio-visual recognizer or used to train a visual recognizer and integrate with the output of an audio recognizer later. The visual feature extraction module will directly affect the recognition accuracy.

In a study about human lip reading [5], Neely found that frontal views of the speaker led to higher recognition rates than profile views. This suggests that frontal pose would be favored over profile poses for the purpose of lip reading by computers, and also helps to explain why most existing lip reading systems make a common assumption that the speaker is in frontal pose. In order to get frontal pose lip image, some researchers use special devices such as a mirror in front of the mouth, and some researchers try to recognize frontal pose and simply drop the non-frontal pose. This makes systems have limited use in tasks such as video transcription (speech recognition of the audio stream for video indexing and retrieval), where speakers are not necessarily in frontal pose.

The literature on dealing with the pose problem in lip reading is fairly limited. Nonetheless, many works have been done in face recognition [7, 4, 1]. Probably the most straightforward way of compensating pose is to represent knowledge of 3D shape with an explicit 3D model such as 3D head wire mesh, by detecting control points on a image, one can warp the image onto the 3D model, then the 3D model is rotated to the new desired pose, and finally the rotated 3D model is projected onto the image plane to get the image of desired pose. This technique is susceptible to feature tracking failure. Another approach is based on example views [1], where images of different poses are captured as examples views. By exploring the relationship between these example views, pose invariant recognition can be achieved. This technique requires careful manipulation of many example views, and the object is assumed to be static, which is not applicable for handling mouth images which are in constant motion during speech.

In broadcast videos, speaker pose changes such as planar translation and rotation can easily be compensated by geometrical transform. For out of plane rotation, as when the speaker rotates his head, simple geometrical transform cannot compensate for the pose.

In this paper, we present a novel method for compensating non-frontal mouth images due to out of plane rotation. By exploiting the general symmetry of human mouth, we prove that from a single profile view, a frontal mouth image can be obtained. Our method is purely working in the image domain, other information such as motion (optical flow), 3D structure of mouth, mouth contour, and segmentation is not needed, thus providing for a general solution for the pose compensation problem in pose invariant lip reading.

The remainder of this paper is organized as follows. In section 2, the general idea behind our method is presented. Then, in section 3, by looking into the imaging geometry, we show that a profile mouth image indeed enables us to recover a frontal pose mouth image. In section 4, the details of our method are described. Our experiments with this method are presented in section 5. Finally, we discuss

Figure 1: Can we recover frontal mouth region (lower middle) by using one non-frontal image (upper left) and the mirror image (upper right)?

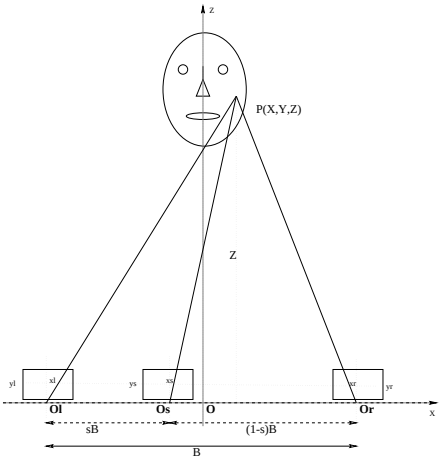

Figure 2: Imaging geometry

conclusions and future work in section 6.

2. THE IDEA

We all may have noticed many amazing video clips in movies and TV commercials depicting the fluid transformation of one digital image into another. This process is called image morphing, which is realized by coupling image warping with color interpolation. Image warping applies 2D geometric transformations on the images to retain geometric alignment between their features, while color interpolation blends their color. If a face of a man can *gradually* be changed to a face of a woman, can we get a frontal pose image from a profile image by *gradually* rotating the head? see Figure 1.

Unfortunately, we only have one profile image, and we don't have a target image to warp onto. Thanks to the *general symmetry* of human face, we can consider the mirror reflection of the profile image as the image seen from the other side. By "morphing" the image, somewhere in the process, we could expect a frontal pose image. That's the general idea of our method, though our method does not use image morphing directly.

In the next section, we can see that by using only one profile view of a mouth image, a frontal mouth image can be recovered based on imaging geometry.

3. IMAGING GEOMETRY

Consider the camera configuration in Figure 2. We can assume the profile image was given by the camera O_l, and the mirror reflection image was given by the camera O_r. Let the origin of the world coordinate system be O, i.e. $O = (0,0,0)$, then $O_l = (-B/2,0,0)$ and $O_r = (B/2,0,0)$. Let $s \in [0..1]$. If we put a virtual camera at $O_s = (-(0.5-s)B,0,0)$, which is sB from O_l, then this virtual camera is $(1-s)B$ from O_r. If $s = 0.5$, the virtual camera is at O, the virtual image is the frontal pose image.

According to the perspective camera model (pin-hole model), a point in 3D space $P(X,Y,Z)$ is projected onto the image plane with optical center at the world origin by the following equation:

$$x_i = f\frac{X}{Z}, \quad y_i = f\frac{Y}{Z}, \tag{1}$$

where f is the focal length of the camera, x_i, y_i is the image plane coordinates. Suppose there is a point on human face, $\vec{P} = (X,Y,Z)$, according to perspective projection camera model, $\vec{P}$ will project onto left camera at the position

$$x_l = f\frac{X+B/2}{Z}, \quad y_l = f\frac{Y}{Z}, \tag{2}$$

while on the right camera, the point will project to

$$x_r = f\frac{X-B/2}{Z}, \quad y_r = f\frac{Y}{Z}. \tag{3}$$

Then on the virtual camera O_s, it will project onto

$$\begin{aligned} x_s &= f\frac{X+(0.5-s)B}{Z} \\ y_s &= f\frac{Y}{Z}. \end{aligned} \tag{4}$$

Representing in terms of x_r and x_l, we have

$$\begin{aligned} x_s &= (1-s)x_l + sx_r \\ y_s &= y_l = y_r \end{aligned} \tag{5}$$

This equation suggest that we can generate any virtual image by a camera on the line between O_l and O_r. Obviously, the front image can be obtained by setting $s = 0.5$, i.e.

$$x_f = \frac{x_l + x_r}{2}, \quad y_f = x_l = x_r. \tag{6}$$

4. THE ALGORITHM

Face and feature locating is a very important problem for a number of applications, including lip reading. Detailed discuss of face and feature locating is beyond the scope of this paper. For face and mouth finding, we use the system reported by Senior [8]. From broadcasting video sequences, the system can output mouth region images with scale, translation, and in-plane rotation compensated. The mouth images are aligned such that the two corners of the mouth lie in a horizontal line. Also, the image sizes are scaled to a certain ratio with respect to the distance between two eyes.

Given such a non-frontal mouth region image, the task is to compensate for rotation in depth using the derivations in the above section. The pose compensation algorithm involves two steps: In order to recover the frontal pose mouth region, first we need to establish correspondences for the pixels in a real view and its mirror reflection. Then the coordinates of matched pixels and their image attributes are used to recover the frontal pose.

4.1. Matching pixels

We use color information as matching attributes. The original color image is in (RGB) format. The differences in RGB values are not necessarily reflecting the color differences in human perception. In order to make the Euclidean distance between pixel attributes reflect the perceptual differences, we use $CIE-L^*a^*b^*$ uniform color system. Due to great variation in the literature, we include the transform as follows [2].

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 2.7690 & 1.7518 & 1.1300 \\ 1.0000 & 4.5907 & 0.0601 \\ 0.0000 & 0.0565 & 5.5943 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (7)$$

$$L^* = 116[Y/Y_0]^{1/3} - 16, Y/Y_0 > 0.01 \quad (8)$$

$$a^* = 500[(X/X_0)^{1/3} - (Y/Y_0)^{1/3}] \quad (9)$$

$$b^* = 200[(Y/Y_0)^{1/3} - (Z/Z_0)^{1/3}] \quad (10)$$

where X_0, Y_0, Z_0 are the XYZ values of the reference white, and for $Y/Y_0 <= 0.01$, the $L^*a^*b^*$ values are set to 0.

Now, let us examine constraints for matching pixels: (1) Epipolar constraint, which requires points on image plane (x_r, y_r), (x_l, y_l), and point on the object P be co-planar. This means a pixel matches to a pixel in the mirror reflection image only if the pixel has the same y coordinates. This becomes obvious from figure 2. (2) Pixel ordering constraint, which requires the pixels matched obey the order from left to right (or right to left). This is a natural constraint, and can be verified from the imaging geometry. See Figure 3.

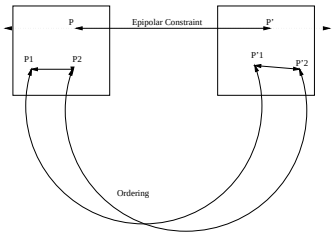

Figure 3: Epipolar and Ordering constraints

Many matching algorithms can be used to get dense pixel correspondence between the given image and its mirror image. We use an algorithm that only requires pixel level features. The pixel matching algorithm takes one non-frontal image as input, without loss of generality, the image can be treated as the image from left camera, then its mirror image can be treated as the image from right camera. Let P_l and P_r be pixels in the left and right images, w the width of correlation window, then for each pixel in left image (x_l, y_l), find its corresponding pixel in the right image

$$x_r = \arg\min_k \sum_{k=x_l-w}^{x_l+w} [L_l^*(x_l, y_l) - L_r^*(k, y_l)]^2 \quad (11)$$
$$+ [a_l^*(x_l, y_l) - a_r^*(k, y_l)]^2 + [b_l^*(x_l, y_l) - b_r^*(k, y_l)]^2$$

where L_r^*, a_r^*, b_r^* stands for the CIELAB color attributes value of P_r, while L_l^*, a_l^*, b_l^* for P_l.

To guarantee robust pixel matching, we use a coarse-to-fine control strategy. In the first round, we use large window size w to estimate the coarse match for each pixel; then in the second round, a smaller window is used to find the precise match as guided by the coarse matches. The ordering constraint is enforced at this step.

4.2. Recovering Frontal Pose Image

Once the pixel correspondences are established, we can calculate the frontal pose image using equation (6), the pixel values are the average of the two matched pixels. Actually, the frontal mouth image in Figure (1) was given by our algorithm.

An interesting property of our method is that an initial pose estimation is not needed. From equation (6), we can see that x_r and x_l can switch position, which means we can treat the real image either as left image or right image while the frontal pose image remains the same.

When lip contour can be reliably detected, the matching algorithm can be very easy. One can use epipolar and ordering constraints to compensate pose by linear interpolation.

5. EXPERIMENTS

We test our method on a set of non-frontal mouth images found from a video sequence. As seen from Figure 4, the frontal pose mouth image is recovered. Because lighting directions are slightly different for images in different poses, the pose compensated frontal pose mouth images are not exactly the same as the true frontal pose mouth image.

To show the effects of pose compensation, we compare the pose compensated images with the true frontal image. In appearance based recognition systems, eigenspace representation is widely used for comparing images. The key is that the Euclidean distance in eigenspace is equivalent to image correlation [10]. For this experiment, the squared Euclidean distance between two images in CIE-LAB space are calculated.

Also, in order to compare mouth images of different poses, the image should be normalized to make them comparable. This can be done by normalizing the distance between two mouth corners and the distance between two eyes, due to the fact that the ratio of these two distances are approximately constant for different poses. In this experiment, we simply align the mouth corners to the same position, and fit images to an 80×100 grid.

The image distance between two images is computed as the squared sum of Euclidean distances in CIE-LAB space on the 80×100 grid. The distance between the true frontal pose image and the non-frontal pose image in Figure 4 (middle left) is $4.6 \times 10^6 \approx 575/pixel$, while for the compensated image in Figure 4 (middle right) the distance is $4.5 \times 10^4 \approx 6/pixel$. The distance between the true frontal image and the non-frontal pose image in Figure 4 (bottom left) is $4.8 \times 10^6 \approx 600/pixel$, while $1.5 \times 10^5 \approx 19/pixel$ for the pose compensated image Figure 4 (bottom right).

If we consider the non-frontal images as noise corrupted frontal image, by doing pose compensation, we actually achieve higher noise-to-signal ratio.

6. CONCLUSIONS

In this paper, we propose a new method to compensate pose changes of mouth images from one profile image. This method works purely in the image domain, and does not require feature detection for compensation. This is especially useful for appearance based lip reading, where mouth images are used for recognition.

When lip shape can be reliably detected, our method can be used to compensate pose simply by linear interpolation. The limitation of our method is that both mouth corners must be visible in the profile image. Further research is needed in order to compensate for arbitrary pose change of the speaker.

Figure 4: True frontal pose image (Upper), non-frontal pose (middle left) and compensated image (middle right), non-frontal pose (bottom left) and compensated image (bottom right)

7. REFERENCES

[1] D. Beymer. *Pose-Invariant face recognition using real and virtual views*. PhD thesis, MIT, 1996.

[2] M. Celenk. a color clustering techniques for image segmentation. *Computer Vision, Graphics, and Image Processing*, 52:145–170, 1990.

[3] T. Chen and R.R. Rao. Audio-visual integration in multimodal communication. *Proceedings of IEEE*, 86(5):837–851, May 1998.

[4] I.A. Essa and A.P. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Trans. on PAMI*, 9(7):757–763, July 1997.

[5] K. Neely. Effect of visual factors on the intelligibility of speech. *J. Acoust. Soc. Amer.*, 28(6):1275–1277, November 1956.

[6] E.D. Petajan. Automatic lipreading to enhance speech recognition. In *IEEE Global Telecommunicaitons Conf.*, pages 265–272, Atlanta, GA, November 1984.

[7] R. Rae and H.J. Ritter. Recognition of human head orientation based on artificial neural networks. *IEEE Trans. on Neural Networks*, 9(2):257–265, March 1998.

[8] A.W. Senior. Face and feature finding for a face recognition system. In *2nd Int'l conf. on Audio- and Visual-based Biometric Person Authentication*, Washington D.C., March 1999.

[9] P.L. Silsbee and A.C. Bovik. Computer lipreading for improved accuracy in automatic speech recognition. *IEEE Trans. on Speech and Audio Processing*, 4(5):337–351, September 1996.

[10] E. Trucco and A. Verri. *introduction techniques for 3-D computer vision*. Prentice Hall, 1998.