6th International Conference on Spoken
Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

LIP REPRESENTATION BY IMAGE ELLIPSE

László Czap
University of Miskolc, Department of Automation
H 3515 Miskolc, Egyetemváros
E-mail: czap@malacka.iit.uni-miskolc.hu

ABSTRACT

Automatic speechreading systems through their use of visual information to support the acoustic signal have been shown to yield better recognition performance than purely acoustic systems, especially when background noise is present. In this paper an answer is sought to the most important questions of speechreading: Which features can represent visual information well? How can they be extracted? Well-known geometric moments are discussed as a means of visual speech representation. Proposed image ellipse axes are shown to be robust and computationally simple features for describing the shape of lips. An intelligibility study was carried out to see which part of the face gives the most support to speechreading. The whole face, mouth or lips were visible dubbed with noisy voice. Visual support to speech perception of the image ellipse model is compared to that of the parts of the natural face.

1. INTRODUCTION

Nowadays we are witnesses of the enormous development of speech recognition by machines. They, however, still perform poorly when background noise is present. It is generally agreed that most visual information is carried by the lips. The inner lips are especially important and some minor improvement comes from the visibility of teeth and tongue. The main difficulty of incorporating visual information into an acoustic recogniser is to find a robust and accurate method for extracting visual speech features. These features should be sensitive to lip movement and invariant for translation, rotation and scale. Several studies have shown that combining visual information with acoustical information can improve the performance of both the human perceiver and the automatic recogniser. All those studies have shown that audio-visual recognition scores are always higher than either the audio or visual ones in all conditions. This is the greatest challenge and the most important objective for bimodal integration.

2. VIDEO PROCESSING

Video processing aims at pre-processing a sequence of images and extract features suitable for recognition. Because the system should be acceptable to the speaker, no special illumination and face position constraint were applied. Conventional make up – red lipstick – is acceptable to female speakers. The variability problem of lip colour estimation is partially solved by red lipstick in this study. To find the region of interest in a full colour image hue of the HSB colour space is a very suitable feature. Skin hue is constant across talkers and even across races despite the fact that lightness can vary significantly. [1, 2] To find the lip the saturation of the same colour space is also robust to illumination conditions. Video pre-processing is related to the feature extraction. A few lip models require finding the lip contours. The system under discussion – as we will see later on – is robust to inaccurate thresholds.

2.1 Feature extraction

Much of the research in speechreading systems is focused on the crucial problem of feature extraction. How can it best transform a sequence of images into feature values that facilitate recognition? The process should be fast, robust, and yield as much information as possible carried by the fewest number of features, removing redundant and linguistically irrelevant information. Whereas there is no one favourite way of representing visual speech there are impressive methods such as dynamic contours [3], manifolds [4], deformable templates [5], and active shape models [6] that have been and are being developed.

Moment functions [7, 8] have a broad spectrum of image analysis such as invariant pattern recognition and object classification. A set of moments computed from a digital image generally represents global characteristics of the image shape and provides lots of information about different types of geometrical features of the image. Functions of geometric moments can be invariant with respect to image plane transformations such as translation, rotation and scale. Geometric moments were the first ones applied to image processing as they are computationally very simple.

Two dimensional geometric moments of *(p+q)*th order are defined as

$$m_{pq} = \iint_{z} x^{p} y^{q} f(x,y)\,dxdy\ , \qquad \text{p,q=0,1,2,3...}$$

where denotes the image region of the x-y plane, which is the domain of the intensity function *f(x,y).*

Geometric moments of different orders represent different spatial characteristics of the image intensity distribution. A set of moments can thus form a global shape descriptor of an image.

Physical interpretation of some geometric moments: By definition, the moment of order zero *(m_{00})* represents the total intensity of an image. First order functions *(m_{01} , m_{10})* provide intensity moments about the *x* and *y* axis, respectively. The intensity centroid *(x0 , y0)* is given by

$$x0 = m10/m00 \; ; \qquad y0 = m01/m00$$

It is convenient to evaluate moments with the origin of the reference system shifted to the intensity centroid of the image. This transformation makes moments independent of the position of the object. Moments computed with respect to the intensity centroid are central moments and defined as

$$m_{pq} = \iint_z (x - x_0)^p (y - y_0)^q f(x, y)\,dxdy \;, \qquad \text{p,q=0,1,2,3....}$$

Second order moments are a measure of variance of the image intensity distribution about the origin. Central moments $_{20}$ and $_{02}$ give the variances about the mean. The covariance measure is given by $_{11}$.

Second order central moments can be thought to be the moments of inertia of the image about a set of reference axes parallel to the image coordinate axes and passing through the intensity centroid. Principal axes of inertia of the image are defined as the set of two orthogonal lines through the image centroid being used as a reference system. Moments of inertia ($_{20}$, $_{02}$) of the image about this reference system are then called principal moments of inertia of the image. If $_{20}$, $_{02}$and $_{11}$ are the second order central moments of an image in its actual image reference frame and if I_1 , I_2 refer to its principal moment of inertia values, then

$$I_1 = \frac{(m_{20} + m_{02}) + [(m_{20} - m_{02})^2 + 4m_{11}^2]^{1/2}}{2}, \qquad \text{and}$$

$$I_2 = \frac{(m_{20} + m_{02}) - [(m_{20} - m_{02})^2 + 4m_{11}^2]^{1/2}}{2}.$$

These equations can be used to define an image ellipse, which has the same moments of inertia as the original image. Lengths a, b of semi-major axis and semi-minor axis of the image ellipse are given by

$$a = 2\,(I_1/ \;_{00})^{1/2} \; ; \qquad b = 2\,(I_2/ \;_{00})^{1/2}.$$

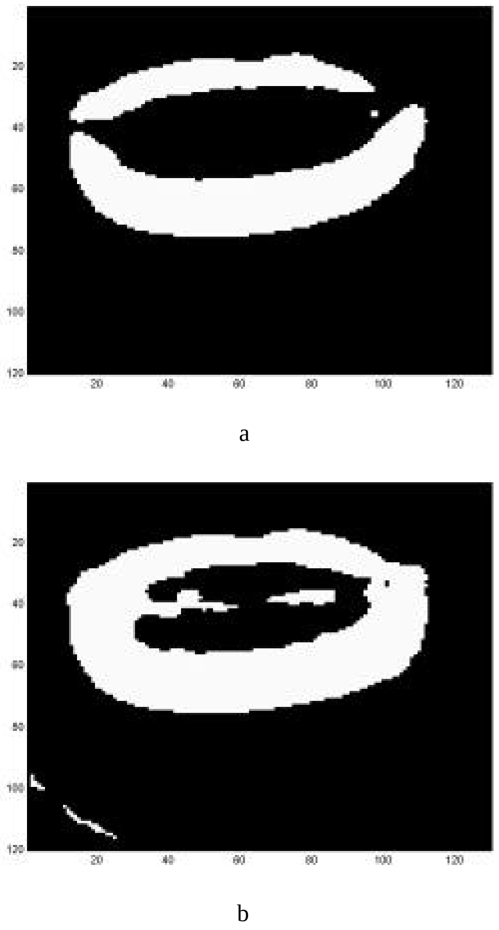
The orientation angle   of one of the principal axes of inertia with the x-axis is given by the following equation:

$$\Theta = \frac{1}{2} tan^{-1}\left(\frac{2\,m_{11}}{m_{20} - m_{02}} \right)$$

The image ellipse also has a uniform intensity value k inside and zero outside preserving the value of the zero order moment defining the intensity factor

$$k = m_{00} \; / (pab) \; .$$

The intensity factor k for the oral cavity carries information on the visibility of the teeth and tongue. An image ellipse can therefore characterise fundamental shape features of an object. The term *s = (I_1+I_2)/m00* is sometimes called shape spreadness, and the term *e = (I_2 – I_1)/(I_2 + I_1)* is called shape elongation.



a

b

Figure 1. Lip shapes extracted with different thresholds providing a=55, b=35 and a=58, b=35 semi-major and semi-minor axes in pixels, respectively.

One of the requirements of feature extraction is to be robust to the variations of lighting and other conditions. The image ellipse model does not require finding the lip contours, and its features are robust to the thresholds of video pre-processing (Figure 1.). Figure 1. a shows a binary image with a high threshold (lip corners are fallen to background) while in Figure 1. b. the threshold is too low, (a part of the tongue and chin contour is considered to belong to the lip). In spite of the difference in the binary image, the extracted geometric features are only slightly different.

Whereas parameterisation of acoustic data is well established, it is not well known which visual features carry the most relevant speech information and which models of the image are most suitable for automatic speechreading.

Figure 2 shows the ellipses of the inner and outer boundaries of the lips applied to the binary image after video processing. The algorithm was developed under MATLAB and was applied to 2685 image frames for experiment discussed in next paragraph.

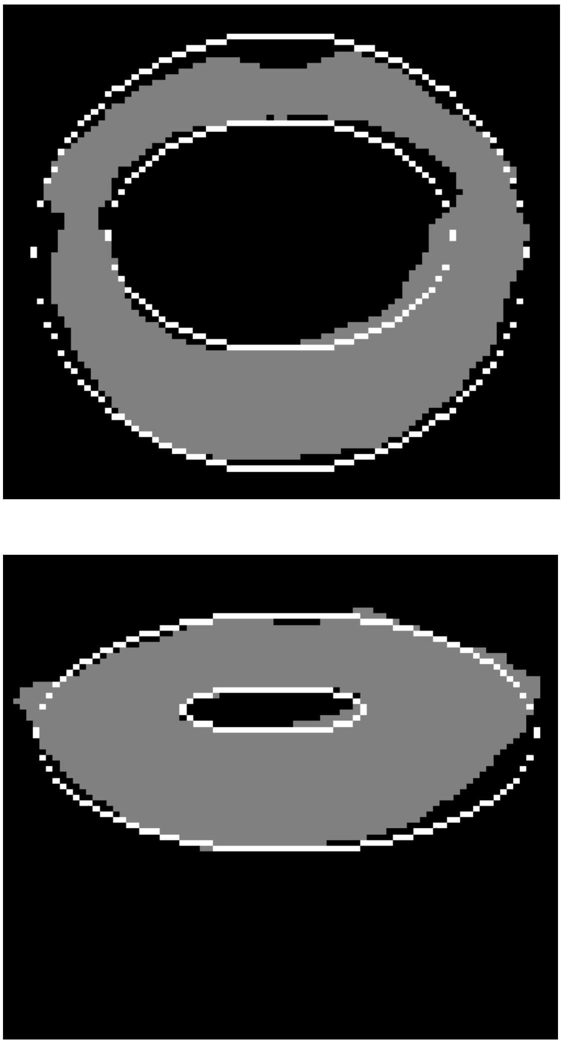


Figure 2. Image ellipses of the inner and outer lip boundaries during the utterance of ‘ a’ (O*) and ‘ u’ (u*) (*SAMPA codes of the vowels).

2.2. An intelligibility study

To find out which parts of the face give more information to a human perceiver an intelligibility study (experiment 1.) was carried out. The corpus was a series of V_1CV_1 words with a consonant between the same vowels. (e.g., ete, ama) and a series of C_1VC_1 words with a vowel between the same consonants. (e.g., bob, tet). The series of consonants and vowels in the middle covered all Hungarian phonemes.

The first stage of our bimodal recognition research aimed at getting information on the visual support of different parts of the face compared to the ellipse model. In the test series the subjects were 78 university students without prior phonetic study. They were asked to listen to the same word twice. Then they wrote down the consonant or the vowel in question. They had limited time for the answer (3 seconds). They were listening to the voice of a series of words each containing the words with only acoustic stimulus then an other series supported with the image sequence of the speaker's lips, then with her mouth (lips, teeth and tongue), then with the whole face. To see the usability of the 2D lip model derived from the image ellipse (filling the area between the two ellipses of Fig. 2 with red), one of the stimuli was an ellipse shown dubbed with noisy voice. Subjects watched the image on the same TV monitor and listened to the voice from a loudspeaker. A clear acoustic stimulus does not need any visual support for recognition. To see the improvement of recognition, the acoustic signal was degraded by additive white noise. The momentary signal to noise ratio was fixed in every 5 milliseconds to -6 dB SNR for consonant and –18dB for vowel experiments. The reason for using momentary SNR was to avoid disturbing consonants more than vowels by an average level of noise. The noise level was calculated to keep the desired signal to noise ratio all the time. There was also an experiment carried out with only the visual stimulus.

Results were obtained after evaluating 11,623 answers. 9,625 of them were to serve consonant recognition and 1,998 of them were used for vowel recognition.

Figures 3, 4, 5 show the results of the intelligibility study of different stimuli.

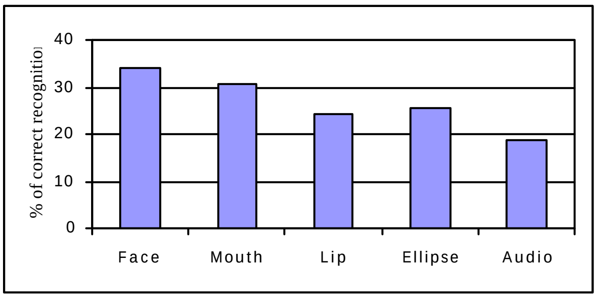


Figure 3.. Rate of the correct answers of the VCV words (consonant recognition), after watching a part of the face (or synthetic lips) and listening to a –6 dB SNR acoustic stimulus.

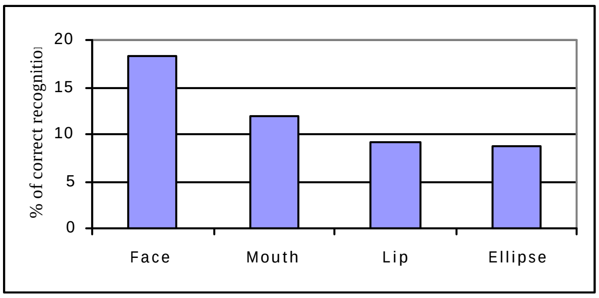


Figure 4. Rate of the correct answers of the VCV words (consonant recognition), by watching a part of the face (or synthetic lips) without any acoustic stimulus.

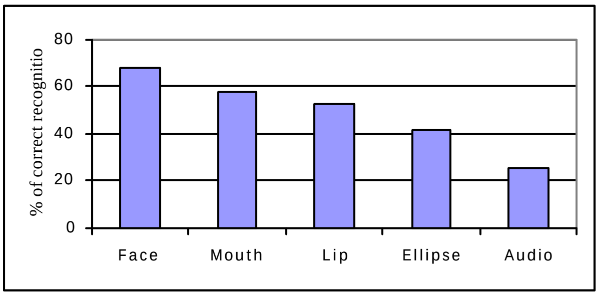


Figure 5. Rate of the correct answers of the CVC words (vowel recognition), through watching a part of the face (or synthetic lips) and listening to a –18 dB SNR acoustic stimulus.

It is not surprising that the whole face contains relevant information to perceive speech. The less part of the face were shown, the fewer words were recognised correctly. Although the 2D ellipse representation of the lips is uncommon, the subjects accepted it. Results of the different stimuli show an error rate of the synthesised stimulus comparable to that of natural lips. These results and the robustness of axis of the image ellipse are encouraging for us to use them as visual features of a speechreading system. To be able to get information on the brightness of the oral cavity (teeth and tongue either visible or not), the intensity factor *k* was added to the *a* and *b* axes of the ellipse of the inner lips. This aims at improving the recognition rate of the ellipse model towards that of a visible mouth.

In the automatic recognition experiment the corpus of speechreading was the manually segmented 120 ms speech of the middle of C_1VC_1 words. (C is selected from b, v, t, l, j and k, while V is selected from a (O), á (a:), e (E), é (e:), i (i), o (o), ö (2), u (u) and ü (y).) The aim of recognition is to identify the vowel, taking the sequence of three images from the manually segmented middle of the vowel (key-frames). Axes *a* and *b* and the intensity factor *k* - calculated for the intensity of oral cavity - were the features of the visual signal. (Visual features spreadness *s*, elongation *e*, and intensity factor *k* were also tried and provided the same results as the previous ones.) Five series of 54 CVC words (six of C by nine of V) were pronounced by a female speaker. Three of the five utterances were used for training and two of them for testing. Training patterns were excluded from testing.

A feed-forward neural network was trained by a back propagation algorithm with visual signals. There were 18 patterns (six surrounding consonants by three utterances) used for training and 12 patterns (two utterances) were used for testing for each of the nine vowels. The 18 training patterns were represented by three neurons in the hidden layer for each vowel. A feed-forward neural network was trained by conjugate gradient back propagation algorithm with Powell-Beale restarts during 2.000 epochs (MATLAB implementation). A recognition rate of 81% was obtained for visual stimuli.

3. CONCLUSIONS

In this paper geometric moments are proposed for lip shape descriptors. An image ellipse can be derived from second order moments that can represent the shape, orientation and position of the lips. An intensity factor represents the visibility of teeth and tongue. Using these features, an 81% recognition rate was reached in a vowel recognition task using visual features by an automatic recogniser. Human perceivers on 75 VCV and 27 CVC words, the subjects judged the consonant or the vowel in the middle of the word evaluated the proposed method. The recognition rate was comparable for natural lips and the synthesised 2D image ellipse lip model. Semi-syllables are the key structures in Hungarian continuous speech recognition [9] and this work is going to be further developed for the Hungarian continuous audio-visual speech recognition.

. REFERENCES

1. Massaro, D.W., Stork, D.G., *Speech recognition and sensory integration,* American Scientist, May-June, 1998.
2. Nankaku, Y., Tokuda, K., and Kitamura, T. *Intensity- and location normalised training for HMM-based visual speech recognition.* In Proceedings of the Eurospeech' 99, Budapest: pp. 1287-1290, 1999.
3. Petajan, E.D. *Automatic lipreading to enhance speech recognition.* In Proceedings of the Global Telecommunications Conference, Atlanta, GA: IEEE Communication Society. pp. 265-272, 1984.
4. Bregler C., and Omohundro, S.M. *Nonlinear image interpolation using manifold learning.* In G. Tesauro, D.S. Touretzky and T.K. Leen (Eds.), Advances in Neural Information Processing Systems, vol. 7, Cambridge, MA: MIT Press. pp. 973-980, 1995.
5. Yuille, A.L., Cohen D.S., and Hallinan, P.W. *Feature Extraction from Faces Using Deformable Templates*. In Proceedings of the Computer Vision and Pattern Recognition. Washington, DC. IEEE Computer Society Press: pp. 104-109, 1989.
6. Luettin, J., Thacker N.A., and Beet, S.W. *Active shape models for visual speech feature extraction.* In D.G. Stork and M.E. Hennecke (Eds.), Speechreading by Humans and Machines, Berlin: Springer-Verlag, pp. 383-390, 1996.
7. Hu, M.K. *Visual pattern recognition by moment invariants.* IRE Transactions on Information Theory, Vol. 8. (1) pp. 179-187, 1962.
8. Mukundan, R., and Ramakrishnan K.R. *Moment functions in image analysis.* Singapore: Word Scientific Press. pp. 11-24, 1998.
9. Vicsi, K., Vigh, A. *Text independent neural network/rule based hybrid, continuous speech recognition*. EUROSPEECH' 95. Madrid: pp. 2201-2204, 1995.