Word Recognition based on Hidden Markov Models with Three-Dimensional Lip Shape Information

Shinji Masuda Norio Tagawa Akihiro Minagawa Tadashi Moriya
Dept. of Electrical Engineering, Tokyo Metropolitan University
1-1 Minami-osawa, Hachioji, Tokyo, 192-0397 Japan
{masuda@elena., tagawa@, akihiro@, moriya@}eei.metro-u.ac.jp

Abstract

We propose a new method of word recognition which is based on Hidden Markov Models (HMMs) and which uses three-dimensional image information. Conventional word recognition methods extract features from two-dimensional image information. However, since such features vary according to the relative position between the camera and face, the recognition rate declines when the speaker is not directly facing the camera [11]. The proposed method extracts features from three-dimensional images generated using high-speed three-dimensional restoration from moiré analysis. The extracted features are modeled by Gaussian distributions and their temporal dependencies by HMMs. We show experimentally, using real sequential images, that the proposed method achieves a higher recognition rate than the subspace method and is not constrained to the relative position between the camera and face.

1 Introduction

Lipreading is to understand speech using visual information of the mouth such as the lip contours of the speaker, shape of tongue or teeth and facial expressions in the absence of sound information. Since a lip reader can understand the context of speech in circumstances in which there is no sound information or when sound information is not effective, if a machine can be developed that is capable of lipreading this would be an important tool for improving the quality of life of people who are hard of hearing and who are not capable of lipreading.

A number of studies have been conducted into this type of man machine interface (MMI) and several algorithms have been proposed. Recently, lipreading has also drawn attention as a way to improve the recognition rate of computer-based speech recognition applications by combining these techniques with lipreading technology.

Lipreading can be divided into two processes: the procedure for extracting the movement of lips as a feature from an image, and the procedure for recognizing the extracted feature. In particular, the former is very important in order for accurate recognition, and several extraction algorithms have been proposed.

The methods for extracting visual speech information from image sequences can be categorized into two groups. One is to use the shape and movement of lip contours as features, and the other is to use all pixels in the neighborhood of the lips as features.

There are several algorithms for extracting the shape of lip contours. Most of these techniques use changes in intensity to extract the shape of lip contours as a feature. In particular, many of these techniques are derived from a minimum energy principle such as Snakes [4]. Although these techniques are comparatively stable, they tend to lack detailed information of the lip contour and it is difficult to define the whole lip contour using the same cost function. In other words, using these techniques, it is difficult to extract the upper lip and lower lip using the same cost function, and it is also difficult to apply a cost function that has the same parameters for different lighting conditions and people. However, since the obtained result is not sufficiently accurate, the lip contour is confirmed using the intensity of pixels in a neighborhood. Luettin et al. have proposed an algorithm for extracting features without Snakes [6]. This technique classifies lip contours by using the shape and intensity of a neighborhood in the same way as the techniques described above.

On the other hand, those methods that use all pixels in the neighborhood of the lips as features, in general, utilize either intensity or optical flow. Li et al. [5] proposed a method for extracting features based on a thresholding process. In this technique, threshold is determined using a gray scale histogram. After the threshold has been determined, features are extracted. This is a simple and high-speed technique. However, it may be unstable with respect to changes

in lighting conditions. Mase et al. [7] proposed a method for extracting features from movement in the neighborhood of the lips calculated as optical flow. Since this method is not based on the shape of lips but movement near the lips, it is hardly affected by differences in people and/or lighting conditions. However, this method has a large computation cost and reduction of this for calculation of optical flow becomes a problem. In addition, it does not use lip shape directly, and does not refer to the position of the lips in an image.

These techniques have been studied based on extracted features from two-dimensional image sequences taken with the speaker directly facing the camera. Under these circumstances, these techniques give good results. However, it is doubtful as to how robust these techniques are when the speaker is allowed to move freely. In other words, these techniques constrain the relative position between the camera and face and this inhibits their practical application. Although experiments were not conducted to evaluate the effects of changing the direction of the face in such papers, it is clear that the experimental results will be adversely affected as demonstrated by Uda et al. [11].

In the present study, we propose an algorithm which overcomes this problem by observing three-dimensional shape of an image using moiré analysis. In moiré analysis, it is possible to extract easily the shape of the lips in real time, and to extract features that do not depend on the position of the face. In addition, by adding depth information of the lips shape to the features the recognition rate is further improved. We demonstrate that the recognition result obtained using three features (width of mouth gap, height of mouth gap and depth of lip contour) is higher than the result obtained from two features (width of mouth gap and height of mouth gap), and also show that the proposed method which uses a hidden Markov model (HMM) achieves a higher recognition rate than the subspace method, a conventional word recognition method.

2 The principle of Moiré Topography

The optical system used for moiré analysis is constructed as shown in Figure1. Light which passes through master grating G_1 with pitch width P_0 from point source S constructs a projective pattern of G_1 on the object surface. F is the distance between S and G_1. This projective pattern of the object is obtained as a transformed grating pattern according to the shape of the object. The transformed pattern is captured by camera through master grating G_2 with pitch P_0, and moiré fringes can then be observed. The phase value of the moiré fringes is calculated using the phase shifting method [2][8][3]. However, shape cannot be decided simply because the moiré fringes do not have constant interval. This problem cannot be solved without knowledge

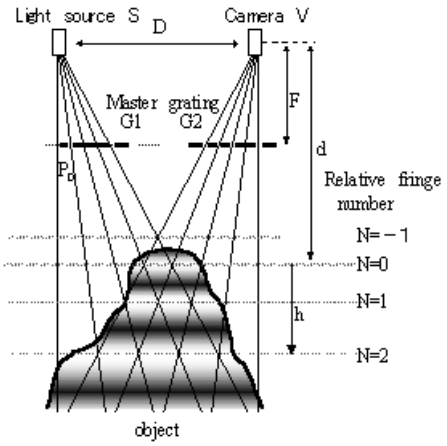


Figure 1. Principle of moiré topography.

of the depth d of the base fringe. To solve this problem, as described by Niikura et al. [8], the depth of the base fringe can be decided by using shading information together with moiré information. Inagaki et al. [3] also reported that it is possible to expand this model to include specular reflection in addition to Lambertian reflection.

3 Feature Extraction

3.1 Extraction of Mouth Gap

In order to lip read using a computer, it is necessary to first extract the lip area of the speaker. In moiré analysis, this is carried out under the hypothesis that the shape of an object is smooth. When the phase value of the moiré fringes is discontinuous, it can be considered to be the lip area because the lip contour is discontinuous. This phase information can be calculated by using the phase shifting method [2][8][3]. According to the phase shifting method, the phase of the moiré fringes can be changed by shifting the observation grating. This is achieved by using an electronic shifting grating. Since the shift value of G_2 is $P_0/2 \times j(j = 0, 1, 2, 3)$, four moiré images can be generated, and in turn four phase images can be generated as shown in Figure2. The discontinuous contour C of the object is regarded as the logical product of the four images thresholding these phase images. Therefore, the gap, which is required for characterization, can be extracted in moiré analysis without the need for additional operations.

3.2 Extraction of Mouth Features

Using the mouth gap extracted as described in the previous subsection, the mouth features are defined as being invariant in relation to face movement. The border line of

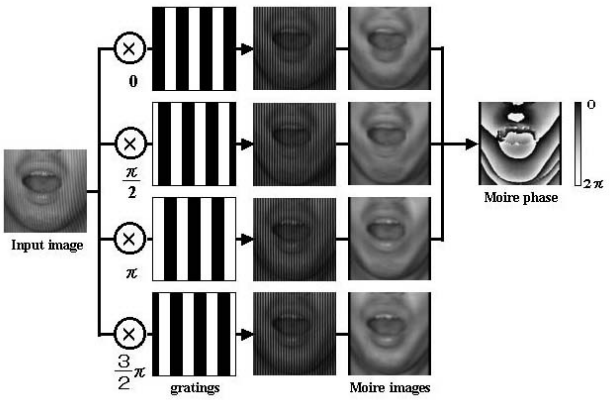


Figure 2. Mouth gap extraction method.

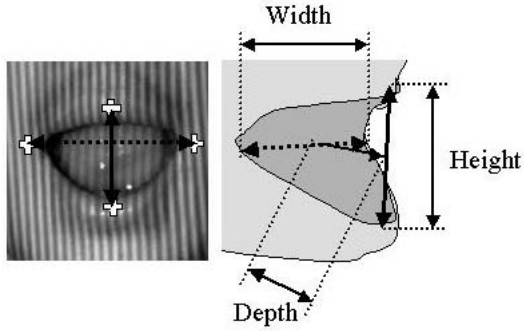


Figure 3. Extracted features from lips image.

the area is considered to be the lip contour. After three-dimensional shape restoration, the height and width of the lips are extracted from the border line and these are considered to be the lip features. The lips features are extracted as follows.

1. Set origin at the center of gravity of the lip contour area. Next, four three-dimensional vectors are set on the contour at initial positions $0, \pi/2, \pi$,and $3\pi/2$ respective to the origin.
2. The plane which has normal vector is defined. This is assumed to be equal to the direction the face is facing.
3. Based on the plane defined in step 2), the deepest two points and shallowest two points are selected.
4. The lip depth is then calculated as shown in Figure 3.
5. If these vectors change between step 2) and step 4) then repeat from 2).

Lip features determined by the above procedure are not sensitive to face movement. In this study, we use extracted features in this way (lip height $a_t$lip width $b_t$lip depth c_t) as temporal vector $\vec{x}_t$.

$$\vec{x}_t = (a_t, b_t, c_t)^T. \quad (1)$$

4 Word Modeling

To model visual speech, we use whole-word HMM (Figure4), which is a standard approach in acoustic small vocabulary recognition systems [10][9][1]. A visual observation of an utterance is represented by a sequence of

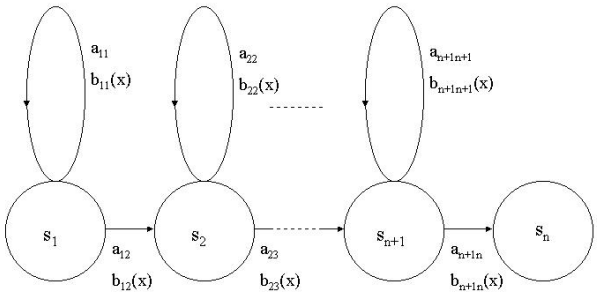


Figure 4. Hidden Markov Model

feature vectors. We assume that the feature vectors follow continuous probability distributions which we model using Gaussians. We trained one HMM for each word class on the corresponding training set for that word. The HMMs only allowed self-loops and sequential transitions between the current state and the next state. The initial state probabilities are set to zero for all states but the first. The remaining parameters are estimated from the extracted model parameters of the training set. The models are further re-estimated using the Baum-Welch procedure, which maximizes the likelihood that the model has generated the observed sequence. Recognition is performed using the forward algorithm which calculates the all state pass sequence for each HMM having generated the observed sequence.

5 Experiment

5.1 Experimental Outline

We evaluated the recognition rate 6 Japanese words (”Chiba”, ”Gifu”, ”Mie”, ”Miyagi”, ”Nagano”, and ”Saga”; 6 Japanese prefectures) spoken by 8 different speakers in order to show the efficiency of the proposed method. Image sequences were sampled at 30 frames per second and the frame size was 256×256 pixels. All images were gray scale with 8bits/pixel. In the present study, the so-called ”cross-variation” technique was applied. The experimental procedure was as follows: 10 of 50 data were used for testing and the remaining 40 data were used for training, so that the number of test data for each word was 50 and the total number of test data was 300.

5.2 Feature Extraction

We extracted features from three-dimensional images obtained by high-speed three-dimensional restoration from moiré analysis. Figure5 shows example images from which features were extracted. The images shown in the figure are two-dimensional. Actually, since lip shape is restored to three dimensions, features can be extracted independent of relative position between the camera and face. Furthermore, since lip shape is restored to three-dimensions, depth information of the lips can be used to represent features. The extracted features as described above are displayed for each word as temporal waveforms and the results for one test are shown in Figure6. Temporal waveforms are different for each word. Therefore, using the differences of these features, we believe that it is possible to build a word recognition system from image information.

As shown in Section 2.1, in the feature extraction method, discontinuous area calculated by Moiré analysis is extracted as the edge, and is connected as a lip contour. However, there is the possibility of mis-extraction of features using this method. Figure7 shows an example of such mis-extraction. The discontinuity at the bottom of the lower lip is very strong, and as such may be incorrectly recognized as the inner area of the mouth. Even though such mis-extraction is rare, it is one of most important problems. We think that this problem can be solved by referring to color images and/or shape information prior to extraction. This problem is a planned topic for future work.

5.3 Comparison of HMM and Subspace Method

We examined the efficiency of word recognition from image information using HMM and compared this to that of the subspace method, one of the conventional word recognition methods. HMMs use 12 states and two-dimensional Gaussian distributions for the training and testing process. In this way, the problem of vector quantification that is one problem of the HMM can be solved. In addition, the discrete data obtained in training is also smoothed. For details of the subspace method used for the experiment refer to [5]. Both the HMM and subspace method use the width and height of mouth gap for recognition. The results of the experiment are shown in Table 1.

Table 1. Difference of recognition rate between HMM and subspace method.

	HMM (%)	Subspace Method (%)
Chiba	94.0	64.0
Gifu	94.0	52.5
Mie	84.0	62.0
Miyagi	96.0	80.0
Nagano	96.0	62.0
Saga	92.0	86.0
average	92.7	66.0

Since the recognition rate of the HMM is better in most cases than that of the subspace method, the HMM was used to construct the word model in the proposed method.

5.4 Efficiency of Depth Information

Next, we examined the efficiency of using depth information of lips for word recognition. We compared the word recognition results obtained using two features (height of mouth gap and width of mouth gap) with the results obtained using three features (height of mouth gap, width of mouth gap and depth of lip contour). For the experiment, HMMs with 12 states were used for word modeling. For word recognition using two features, the HMMs were constructed assuming the observation probabilities to be a two-dimensional Gaussian distribution. For word recognition using three features, the HMMs were constructed assuming the observation probabilities to be a three-dimensional Gaussian distribution. The results are shown in Table 2.

Overall, the recognition rate obtained using three features (width of mouth gap, height of mouth gap and depth of lip contour) is better than that obtained using two features (width of mouth gap and height of mouth gap). Although very little difference is observed in the temporal waveforms of the depth of lip contour between each test word as seen in Figure6, depth information serves to improve the recognition rate in some cases. In other words, depth information reduces the possibility of false recognition. For example, the only recognizable features of the Japanese word ”Mie” are ”mi” and ”e” pronounced [mi:] and [e], respectively.

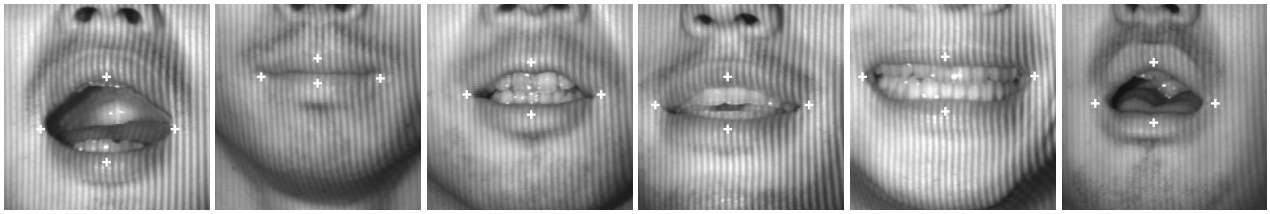


Figure 5. Extracted features.

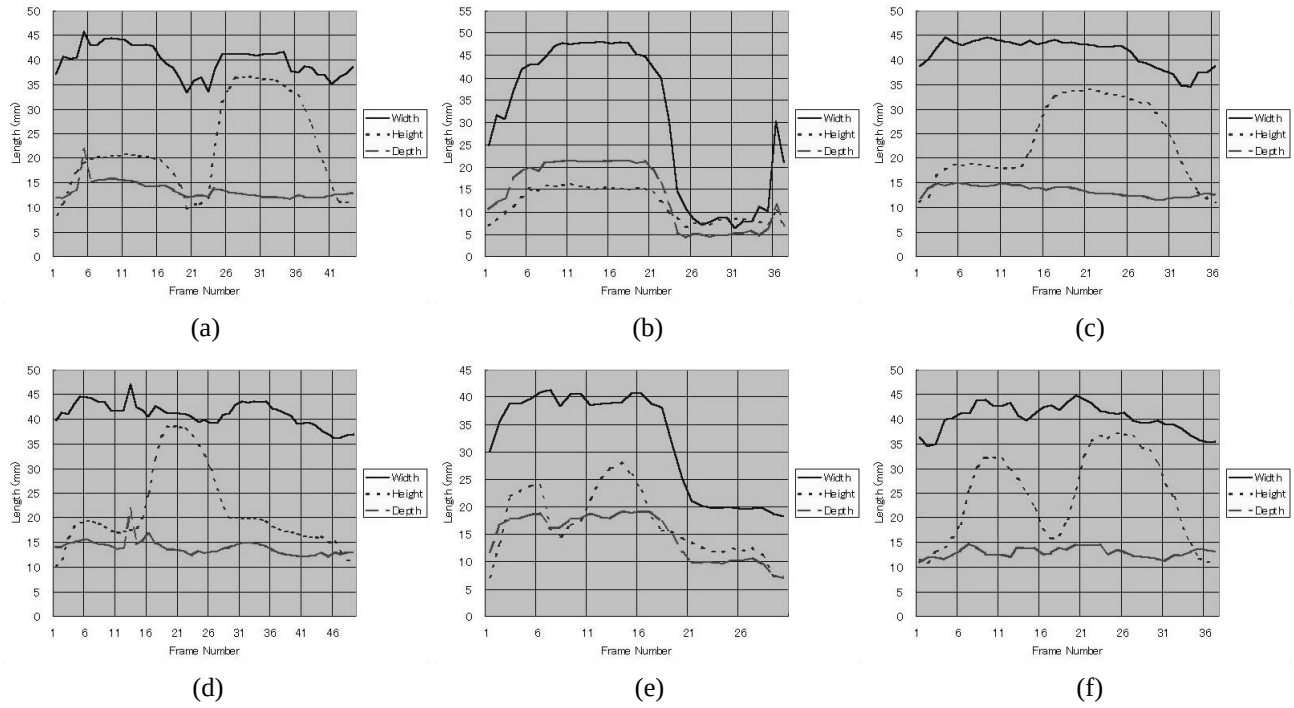


Figure 6. Temporal waveforms of (a) Chiba, (b) Gifu, (c) Mie, (d) Miyagi, (e) Nagano, and (f) Saga.

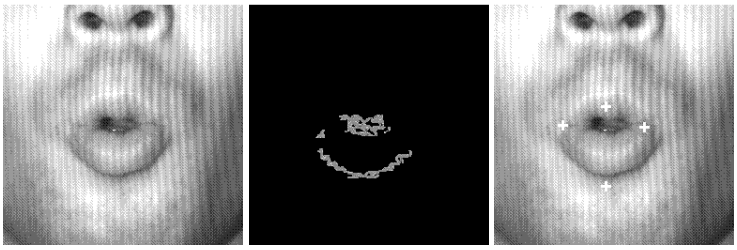
Since ”a” and ”e” pronounced [ʌ] and [e], respectively, in Japanese are similar in terms of height of mouth gap and width of mouth gap, false recognition increases when using only two features. However using depth information, the difference between ”a” and ”e” is more apparent and the recognition rate improves. Therefore, the depth of the lip contour is effective for word recognition.

6 Conclusions

In the present study, we have proposed a new word recognition method which restores three-dimensional lip shape from moiré analysis. Since features are extracted from three-dimensional lip shape, the method is not dependent on the relative position between the camera and face. Furthermore, this method uses the depth of the lip contour in addition to the height and width of the mouth gap and this has been shown to improve the accuracy of word recognition. In addition, the recognition rate results obtained using the HMM of the proposed method were compared with those of the subspace method, a conventional word recognition method. It was found that, overall, the HMM resulted in a higher recognition rate than that of the subspace method. In the future, we plan to extend this study and evaluate the recognition rate obtained using the proposed system for a wider variety of test words.

References

[1] T. Hanazawa, T. Kawabata, and K. Shikano. Phoneme recognition using hidden markov models. *ATR Technical Report*, TR-I-0147, 2 1990.



(a) (b) (c)

Figure 7. Example of mis-extraction, (a) input, (b) discontinuous points, and (c) extracted feature.

Table 2. Difference of recognition rate between 2- or 3-dimensional HMM.

	HMM2-D (%)	HMM3-D (%)
Chiba	94.0	98.0
Gifu	94.0	94.0
Mie	84.0	94.0
Miyagi	96.0	96.0
Nagano	96.0	100.0
Saga	92.0	98.0
average	92.7	96.7

[2] T. Hanazawa, T. Kawabata, and K. Shikano. *Measurement Methods using Moiré Techniques*. Corona Publishing Tokyo, 1996 (in Japanese).

[3] A. Inagaki, N. Tagawa, A. Minagawa, and T. Moriya. Computation of shape and reflectance of 3-d object using moiré phase and reflection model. *Proc. of IEEE International Conference on Image Processing*, (10), 2001(Accepted).

[4] M. Kass, A. Witkin, and D. Terzopoulos. Snakes : active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988.

[5] M. Li, I. Yamasaki, Y. Kurohata, and H. Ogawa. Automatic lipreading by subspace method. *Technical Report of IEICE of Japan*, PRMU97-105, 9 1997 (in Japanese).

[6] J. Luettin and N. A. Thacker. Speech reading using probabilistic models. *Computer Vision and Image Understanding*, 65(2):163–178, 2 1997.

[7] K. Mase and A. Pentland. Automatic lipreading by optical flow analysis. *Transactions of IEICE of Japan*, J73-D-II(6), 1990 (in Japanese).

[8] K. Niikura, A. Minagawa, N. Tagawa, and T. Moriya. Determination of the absolute number of moiré fringes using shading information. *Proc. of 4th International Conference on Advances in Pattern Recognition and Digital Techniques*, pages 154–159, 12 1999.

[9] M. Okochi. Speech recognition based on hidden markov models. *The Journal of the Acoustical Society of Japan*, 42(12):936–941, 1986 (in Japanese).

[10] K. Shikano, S. Nakamura, and S. Ise. *Speech digital signal processing*. Syokodo Tokyo, 1997(in Japanese).

[11] K. Uda, N. Tagawa, A. Minagawa, and T. Moriya. Effectiveness evaluation of speech features including 3-d information for lipreading. *Proc. The 2000 Inf. and Syst. Society Conf. of IEICE of Japan*, D-12-54, 2000 (in Japanese).