Automatic Lipreading with Limited Training Data

Author
Wang, S., Lau, W., C Liew, A., Leung, S.

Published
2006

Conference Title
The 18th International Conference on Pattern Recognition

DOI
https://doi.org/10.1109/ICPR.2006.301

Copyright Statement
Copyright 2006 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Downloaded from
http://hdl.handle.net/10072/24387
AUTOMATIC LIPREADING WITH LIMITED TRAINING DATA

S.L.Wang†, W.H.Lau*, A. W. C. Liew# and S.H.Leung*

†School of Info. Security Engg.  Shanghai Jiaotong University  Shanghai, CHINA
*Dept. of Electronic Engg.  City University of Hong Kong  Kowloon, HONG KONG
#Dept. of Comp. Sci. & Engg.  Chinese University of Hong Kong  Shatin, HONG KONG

ABSTRACT

Speech recognition solely based on visual information such as the lip shape and its movement is referred to as lipreading. This paper presents an automatic lipreading technique for speaker dependent (SD) and speaker independent (SI) speech recognition tasks. Since the visual features are derived according to the frame rate of the video sequence, spline representation is then employed to translate the discrete-time sampled visual features into continuous domain. The spline coefficients in the same word class are constrained to have similar expression and can be estimated from the training data by the EM algorithm. In addition, an adaptive multi-model approach is proposed to overcome the variation caused by different speaking style in speaker-independent recognition task. The experiments are carried out to recognize the ten English digits and an accuracy of 96% for speaker dependent recognition and 88% for speaker independent recognition have been achieved, which shows the superiority of our approach compared with other classifiers investigated.

1. INTRODUCTION

Audio-visual speech recognition has attracted significant interest in recent years since the visual information extracted from the lip movement has been shown to improve the performance of automatic speech recognition (ASR) system especially in noisy environments [1,2]. However, the performance of automatic lipreading, i.e., speech recognition from visual information alone, is far from satisfactory. There are two important issues related to this topic: (i) to extract the speech-relevant lip features accurately and robustly from a lip image sequence; (ii) to choose the most appropriate classifier for the recognition task.

Two kinds of features: pixel-based (modeless) [1] and model-based features [1,2] are widely used in the automatic lipreading. Pixel-based features are usually derived directly from the image after some form of filtering. However, they are usually of high dimension and high redundancy. In addition, it is also difficult for the classifier to extract speech relevant information from these features while ignoring the effects of scaling, rotation, translation and illumination. Model-based features are of low dimensionality and invariant to the varying factors mentioned above. With the lip segmentation and lip contour extraction techniques we proposed recently [3,4], certain model-based features can be derived accurately.

With regard to the classifier selection, many researchers resort to the hidden Markov model (HMM) [1,2] since it performs well in audio speech recognition and has advantage in audio-visual data fusion. However, HMM is a relatively complex statistical model, using limited training data is insufficient to optimize the model parameters. In our approach, we build a statistical word model based on the training data, and perform classification using the maximum likelihood (ML) principle. To handle the variable frame length of the word sequence, we represent each visual feature using a spline curve, thus transforming the discrete time measurement to the continuous domain. The spline curves of the same word class are assumed to have a common mean and a zero-mean random component and the EM algorithm is used to derive these quantities. The final classification is performed by selecting a word model for which the visual feature sequence has the largest ML probability.

For speaker independent recognition task, large variations of visual features caused by different speaking style of different individuals are inevitable. Hence, one word model for each digit may not be able to provide accurate recognition due to such large variations. In this paper, we will demonstrate the use of an adaptive multi-model approach to overcome such variations due to different talking habit.

The paper is organized as follows. In section 2, the derivation of the visual feature set is described. Section 3 presents the spline model for each visual feature and the classification methods for speaker-dependent and speak-independent visual speech recognition tasks. Experimental results and comparison with other widely used techniques are shown in section 4. Finally, conclusion is given in section 5.

2. VISUAL FEATURE EXTRACTION

In our previous work, a lip segmentation technique [3] and a lip contour extraction method [4] have been proposed.
to obtain the lip outer contour accurately from the lip images. Then the following three kinds of lip features were extracted from the lip contour for recognition.

i) Geometric features: width and height of the lip. In order to reduce the variations caused by different distances between the speaker and the camera, these features are normalized against those of the first image in the sequence.

ii) Shape descriptors: ASM features

Active Shape Model (ASM) is used in our lip modeling approach [4] and the coordinates of the contour points of an arbitrary shape $\mathbf{x}$ can be approximated by (1).

$$\mathbf{x} = \mathbf{x} + \mathbf{Pb}$$

where $\mathbf{x}$ is the mean shape, $\mathbf{P}$ is the matrix of eigenvectors of the covariance matrix and $\mathbf{b}$ is the weight vector for each eigenvector. It should be noted that only the first few eigenvectors corresponding to the largest eigenvalues are sufficient for modeling the shape variation. In our approach, we use the first three weight values.

iii) Inner mouth features: teeth and mouth opening 

By using the teeth detection technique in [6], the total area of the teeth region is calculated and represented by $\text{area}$ which is normalized against the entire mouth area enclosed by the outer lip contour. The inner mouth opening $m_{\text{open}}$ is another parameter considered for recognition and it is obtained by analyzing the luminance change and the teeth information along the middle axis of the mouth. The opening is also normalized against the height of the mouth region.

![Fig.1 (a) The original lip image, (b) teeth region and inner mouth opening (a) (the white pixels represent the teeth)](image)

**3. A SPLINE MODEL FOR VISUAL FEATURE CLASSIFICATION**

**3.1 Spline representation of the visual feature sequence**

Though most of the visual features are continuous in nature, their observations are made on discrete time points governed by the sampling rate of the lip image sequence (usually 25 or 30 frames per second). In our approach, B-spline curves are used to transform the discrete observations to a continuous domain. Instead of fitting any observations to a continuous domain, instead of fitting any

The observation sequences of each visual feature in the same word class are modeled by

$$Y_{ic} = S_{ic}(\mu_c + \gamma_{ic}) + \epsilon_{ic}, \quad i = 1, 2, ..., m_c, \ c = 1, 2, ..., C$$

where $Y_{ic}$ denotes the observation values for the $i$th sequence in class $c$ at time points $t_1, t_2, ..., t_n$ ($n$ is the number of sampling points for the sequence); $S_{ic}$ is the spline basis functions evaluated at those time points; $\mu_c$ is the common mean of the spline coefficients in class $c$; $m_c$ is the number of training sequences in class $c$; $C$ is the number of word classes for classification; $\gamma_{ic}$ is a zero-mean random component representing the variation among all the observation sequences in class $c$ and finally $\epsilon_{ic}$ is the vector noise terms. Note that the control points of the splines are uniformly spaced covering the entire duration of the utterance. If $Y_{ic}$ and $\epsilon_{ic}$ are further assumed to be Gaussian distributed, i.e., $Y_{ic} \sim N(0, \Gamma_{ic}), \epsilon_{ic} \sim N(0, \sigma^2 I)$, $Y_{ic}$ is then also Gaussian distributed with $Y_{ic} \sim N(S_{ic} \mu_c, \Sigma)$ where $\Sigma = \sigma^2 I + S_{ic} \Gamma_{ic} S_{ic}^T$. Hence the joint distribution of the observed curves can be written as

$$\prod_{c=1}^{C} \prod_{i=1}^{m_c} \frac{1}{\sqrt{2\pi \Sigma_{ic}}} \exp\left(-\frac{1}{2} (Y_{ic} - S_{ic} \mu_c)^T \Sigma_{ic}^{-1} (Y_{ic} - S_{ic} \mu_c)\right)$$

Directly estimating $\mu_c$, $\sigma$ and $\Gamma_c$ by maximizing the joint probability in (3) is a difficult non-convex optimization problem [5]. However, if the $\gamma_{ic}$ values are available, the joint probability in (3) can be simplified to

$$\prod_{c=1}^{C} \prod_{i=1}^{m_c} \frac{1}{\sqrt{2\pi \Sigma_{ic}}} \exp\left(-\frac{1}{2\sigma^2} (Y_{ic} - S_{ic} \mu_c)^T (Y_{ic} - S_{ic} \mu_c)\right)$$

where $k$ is the dimension of the spline coefficients. The EM algorithm can be employed to maximize the likelihood in (4). In the E step, the expected value of $\gamma_{ic}$ is calculated with $\mu_c$, $\sigma$ and $\Gamma_c$ fixed. In the M step, $\mu_c$, $\sigma$ and $\Gamma_c$ can be estimated by maximizing the joint likelihood in (4) with known $\gamma_{ic}$. The updating formulae of the parameters are shown in the Appendix and readers can refer to [5] for detail derivations.

**3.2 Speaker dependent classification method**

With the observation sequence of the $j$th visual feature in class $c$ expressed by a Gaussian model with mean $\mu_{jc}$ and covariance matrix $\Sigma_{jc}$, the classification algorithm for speaker dependent tests can be run as follows:

i) From the observation values obtained from the training data, the $\mu_{jc}$ and $\Gamma_{jc}$ are estimated for each visual feature $\nu_j$ and each class $c$ using the EM algorithm. The variance of the noise term $\sigma$ is also calculated.

ii) Let $Y_{ic}$ denotes the observation sequence of the $j$th visual feature and $f$ be the number of visual features. The testing sequence can be expressed by
\[{Y_{v_1}, Y_{v_2}, \ldots, Y_{v_f}}\} \). The probability of the testing sequence belonging to class \(c\) can be formulated as
\[
p_c = \prod_{j=1}^{f} p(Y_{v_j} | \mu_{jc}, \sigma_{jc})^{w_j}, \]
where \(w_j\) is the weight factor of feature \(v_j\) and it is set to be the same value \((w_j = 1/f)\) in our experiment.

ii) The correct word model for the testing sequence is classified to \(\arg \max_c p_c\).

It should be noted that each visual feature is assumed independent to each other in the proposed approach. Although this assumption may not be true for all the features, it greatly reduces the computational complexity and provides satisfactory results at the same time.

### 3.3 Speaker independent classification method

For speaker independent tests, large variation caused by different speaking style and habit of individual speaker is inevitable. One single model for each digit cannot deliver satisfactory results in this case. In order to solve the problem, a multi-model approach is proposed and the details are presented in the following:

i) For any sequence \({Y_{v_1}, Y_{v_2}, \ldots, Y_{v_f}}\) in the training data, calculate the spline coefficients for each visual feature sequence by least-squared fitting method, i.e.,
\[
l_{v_j} = (S^T S)^{-1} S^T Y_{v_j} \quad j = 1, 2, \ldots, f\]. Then the sequence is represented by a spline coefficient vector
\[
l = [l_{v_1}^T, l_{v_2}^T, \ldots, l_{v_f}^T]^T\]. Initialize the number of groups with \(M=1\).

ii) Partition the spline coefficient vectors of training data set into \(M\) groups by \(k\)-means algorithm. Store the group partition information.

iii) Compute the Euclidean distance between each coefficient vector and the centroid of the group it belongs to and calculate the standard deviations of all the groups.

iv) Repeat step (2) with \(M = M + 1\) until the maximum standard deviation is greater than that of last iteration or any group contains less than 3 components.

v) The appropriate number of groups is set to \(M - 1\) and the group information for \(M - 1\) is restored.

vi) Unlike the single speaker case, \(M - 1\) models have been trained for the same digit based on the group partition (\(M\) could be different for different word class). The testing sequence is classified to digit \(c\) if the output probability of any of the models belonging to digit \(c\) is the highest among all the models.

Note that similar approaches based on the “multi-model” idea can also be properly modified and applied to RDA and HMM.

### 4. EXPERIMENTAL RESULTS

In order to evaluate the performance of the proposed automatic lipreading algorithm, a database consists of 10 isolated English digits (0 to 9) uttered by eight different individuals, 6 males and 2 females, has been established. Each speaker was asked to repeat every digit for twenty times and each digit utterance contains 30 lip images with size 110 by 90 and there are about 50,000 lip images in the database.

#### 4.1 Speaker-dependent recognition tests

Tables 1 to 3 show the recognition performance using our approach compared with two widely used classification methods: Hidden Markov Model (HMM) and Regularized Discriminant Analysis (RDA) [7]. The HMM is defined as a left to right, six states, continuous density with diagonal covariance matrix Gaussian modes associated with each state. The Baum-Welch algorithm following the Maximum Likelihood (ML) criterion have been used for training the HMM, and the Viterbi algorithm for recognition. The regularization parameter \(\gamma\) [7] for the RDA is set to 0.5 and the order of RDA is two. And the B-spline used in our approach is a cubic one with 5 knots. These parameter settings provide better results than many other choices in each classification technique. One hundred random tests have been performed for each test and the average recognition rate is taken as the accuracy.

<table>
<thead>
<tr>
<th>Classification technique</th>
<th>Number of training samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM</td>
<td>5</td>
</tr>
<tr>
<td>RDA</td>
<td>76.28</td>
</tr>
<tr>
<td>Our approach</td>
<td>85.03</td>
</tr>
</tbody>
</table>

Table 1. The speaker dependent recognition accuracy in % obtained using the geometric features: width and height of the lip.

<table>
<thead>
<tr>
<th>Classification technique</th>
<th>Number of training samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM</td>
<td>5</td>
</tr>
<tr>
<td>RDA</td>
<td>86.30</td>
</tr>
<tr>
<td>Our approach</td>
<td>91.98</td>
</tr>
</tbody>
</table>

Table 2. The speaker dependent recognition accuracy in % obtained using the geometric features and the ASM features b.

<table>
<thead>
<tr>
<th>Classification technique</th>
<th>Number of training samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM</td>
<td>5</td>
</tr>
<tr>
<td>RDA</td>
<td>88.77</td>
</tr>
<tr>
<td>Our approach</td>
<td>94.50</td>
</tr>
</tbody>
</table>

Table 3. The speaker dependent recognition accuracy in % obtained using all the visual features introduced in Section 2.

It can be seen from the results that the proposed algorithm always outperforms the other two methods. HMM does not perform well because the number of training samples is not large enough to train a sufficient model and that in turn leads to overfitting. RDA does not achieve satisfactory results either because it describes the
relationship among the observations of a sequence by correlation. The proposed algorithm uses a spline curve to represent the static and dynamic information of the visual feature so that higher order relationship is inherently embedded in the spline coefficients. The proposed algorithm could also gives reliable results with just 5 training samples. In contrast, some models using HMM cannot even be initialized with such a small number of training samples.

4.3 Speaker-independent recognition tests

Table 4 shows the speaker independent recognition accuracy obtained by the three methods investigated. For each digit, the training data consists of half of the total samples for all speakers and the rest is for testing. One hundred random tests have also been performed to reduce the effect of using a specified training data.

<table>
<thead>
<tr>
<th>Classification method</th>
<th>Single model</th>
<th>Multiple models</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM</td>
<td>75.87</td>
<td>86.32</td>
</tr>
<tr>
<td>RDA</td>
<td>62.37</td>
<td>72.45</td>
</tr>
<tr>
<td>Our approach</td>
<td>73.30</td>
<td>87.07</td>
</tr>
</tbody>
</table>

Table 4. The speaker independent recognition accuracy in % using all the visual features introduced in Section 2.

From the table, the performance of HMM is similar to that of the proposed method with single model for each digit and it is considerably better than that of RDA. This is because large variation due to different speaking style and habit of various speakers is unavoidable in speaker independent tests. A simple Gaussian distribution of RDA is insufficient to model the variation whereas the parameters in HMM are well estimated by adequate training samples. It is also observed from the experimental results that the accuracy for speaker independent recognition tests can be greatly improved by the multi-model approach for all the three classification methods. In such approach, one can use more models to represent the samples if large variations are present in the database due to different speaking styles. Each model is then focused on a particular kind of speaking style and the speech-irrelevant information, the different speaking style of various speakers uttering the same word is therefore discarded.

5. CONCLUSIONS

This paper presents an isolated word recognition technique solely based on the visual features. The visual speech features are composed of the geometric width and height of the lip, the ASM shape descriptors and the inner mouth features. A spline curve has been employed to model the static and dynamic information for each visual feature in the same word class and the spline coefficients are estimated from all the training samples within the class by the EM algorithm to avoid overfitting. A multi-model approach has also been proposed to overcome the variation caused by different speaking habit and style for the speaker independent visual speech recognition tests. Experimental results demonstrated that the proposed algorithm provides accurate recognition results for both speaker dependent and speaker independent tests even when only limited training data is available.

6. APPENDIX

Taking the logarithmic of (4), we have:

\[
\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{m} \left( \frac{1}{\sigma^2} (Y_{ij} - S_{ij}(\mu_i + \gamma_{ij}))^T (Y_{ij} - S_{ij}(\mu_i + \gamma_{ij})) + \gamma_{ij}^2 \Gamma_{ij}^{-1} \gamma_{ij} \right) + n_i \log \sigma^2 + (n_i + k) \log(2\pi) + \log \Gamma_i + \log \mu_i 
\]

By discarding the constant term, the maximization (A1) is equivalent to minimize

\[
\sum_{i=1}^{n} \sum_{j=1}^{m} \left( \frac{1}{\sigma^2} (Y_{ij} - S_{ij}(\mu_i + \gamma_{ij}))^T (Y_{ij} - S_{ij}(\mu_i + \gamma_{ij})) + \gamma_{ij}^2 \Gamma_{ij}^{-1} \gamma_{ij} + n_i \log \sigma^2 + \log \Gamma_i + \log \mu_i \right) 
\]

In the E step, the expect values for \( \gamma_{ijc} \) can be computed by

\[
\gamma_{ijc} = (\sigma^{-2} \Gamma_{ij}^{-1} + S_{ij}^T S_{ij})^{-1} S_{ij}^T (Y_{ij} - S_{ij} \mu_{ij}) 
\]

In the M step, \( \sigma^2 \), \( \mu_{ij} \) and \( \Gamma_{ij} \) are updated to maximize the joint likelihood w.r.t. the expected values of \( \gamma_{ijc} \):

\[
\sigma^2 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} [Y_{ij} - S_{ij}(\mu_{ij} + \gamma_{ij})]^T (Y_{ij} - S_{ij}(\mu_{ij} + \gamma_{ij})) + \text{trace}(S_{ij}^T [\Gamma_{ij}^{-1} + S_{ij}^T S_{ij}/\sigma^2]) S_{ij}] + \sum_{i=1}^{n} \sum_{j=1}^{m} n_i \log \sigma^2 + \log \Gamma_i + \log \mu_i}{n} 
\]

7. REFERENCES