J. Vis. Commun. Image R. 30 (2015) 219–233

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Geometrical-based lip-reading using template probabilistic multi-dimension dynamic time warping ☆

M.Z. Ibrahim [a,*], D.J. Mulvaney [b]

[a] Faculty of Electrical and Electronics Engineering, University Malaysia Pahang, 26300 Pahang, Malaysia
[b] School of Electronic, Electrical and Systems Engineering, Loughborough University, LE11 3TU, United Kingdom

ARTICLE INFO

Article history:
Received 12 March 2013
Accepted 27 April 2015
Available online 5 May 2015

Keywords:
Lip reading
Lip geometry
Mouth detection
Skin segmentation
Convex hull
Multi dimension dynamic time warping
Template probabilistic
OpenCV

ABSTRACT

By identifying lip movements and characterizing their associations with speech sounds, the performance of speech recognition systems can be improved, particularly when operating in noisy environments. In this paper, we present a geometrical-based automatic lip reading system that extracts the lip region from images using conventional techniques, but the contour itself is extracted using a novel application of a combination of border following and convex hull approaches. Classification is carried out using an enhanced dynamic time warping technique that has the ability to operate in multiple dimensions and a template probability technique that is able to compensate for differences in the way words are uttered in the training set. The performance of the new system has been assessed in recognition of the English digits 0 to 9 as available in the CUAVE database. The experimental results obtained from the new approach compared favorably with those of existing lip reading approaches, achieving a word recognition accuracy of up to 71% with the visual information being obtained from estimates of lip height, width and their ratio.

1. Introduction

Automatic speech recognition (ASR) systems are starting to play an important role in human computer interfaces (HCI); for example Siri, marketed as the intelligent personal assistant for the iPhone 4S, is able to respond to spoken user requests [1]. In controlled environments, modern ASR systems are capable of producing reliable results, but in many real-world situations the intrusion of acoustic noise adversely affects performance and the resulting recognition rates are often adversely affected [2]. As many potential ASR users will use mobile devices in noisy environments such as vehicles, offices, airport terminals and train stations, solutions that provide reliable operation at high ambient noise levels will become increasingly important.

Humans are often able to compensate for noise degradation and uncertainty in speech information by augmenting audio with visual information to aid the recognition task [3]. People with hearing impairments may have a reduced ability to receive information in the audio domain, so they rely more heavily on the visual domain for speech recognition. The mechanism employed is often termed either 'lip reading' or 'speechreading'. Lip reading is the ability or skill to understand speech through information gleaned from the lower part of face, typically by following lip, tongue and jaw movement patterns. Speechreading may include lip reading information, but may provide additional understanding of the speech by interpreting whole face expressions, gestures and body language [4–6], as well as environmental conditions, such as specific characteristics of the speaker and the time and physical location at which the conversation took place [7].

When integrating lip reading or speechreading into an ASR system, one of the fundamental issues to address is the selection of the visual features that will be the most advantageous in enhancing recognition performance. Research centers on two different types of feature, namely appearance-based and shape-based features. Appearance-based features are used to model characteristic within the mouth region, typically capturing information related to spatial frequencies, whereas shape-based features extract geometrical measurements normally relating to the shape of the lips. In the work presented in this paper, we have chosen to concentrate on extracting shape-based features from the lip region, as these are likely to contain the visual information most closely related to the spoken sounds, and the lip movements will be highly correlated with the speech sounds themselves, making integration with speech features more straightforward.

☆ This paper has been recommended for acceptance by Yehoshua Zeevi.

* Corresponding author.
E-mail addresses: zamri@ump.edu.my (M.Z. Ibrahim), d.j.mulvaney@lboro.ac.uk (D.J. Mulvaney).

http://dx.doi.org/10.1016/j.jvcir.2015.04.013
1047-3203/

Several lip reading approaches have been described in the literature. In [8], the results of visual ASR experiments involving the use of the IBM ViaVoice database were presented in their comparison of four types of visual features, namely discrete cosine transform (DCT) [9], discrete wavelet transform (DWT) [10], principal components analysis (PCA) [11], and active appearance models (AAM) [12]. A solution using hidden Markov models (HMMs) [13] as the classifier found that DCT based visual features were the most promising for the recognition task. In [14], both appearance and shape based visual features were obtained using PCA applied to facial animation parameters (FAPs) [15] obtained from outer and inner lip contours that in turn were found by tracking using a combination of a gradient vector field (GVF) [16] and a parabolic template. The experiments show that under challenging visual conditions (involving changes in head pose and lighting conditions), the lip reading performance of appearance-based visual features suffered. It was also shown that the features obtained from inner-lip FAPs did not provide as much information useful for lip reading as did those obtained from the outer-lip FAPs.

In [5], hue and canny edge detection [17] were used to segment the lip region, extracting shape-based features including lower and upper mouth width, mouth opening height and distance between horizontal lip line and upper lip. These features were used in experiments to recognize 78 isolated words using an HMM classifier. Ten subjects from the Carnegie Mellon University database [18] were used to evaluate the performance of the system, with a best classification performance of 46% accuracy being attained when all the geometrical information and delta features were included and when operating in speaker dependent mode. The performance was found to fall to 21% in the speaker independent case.

In [19], the lip region was located using a Bayesian classifier [20] that held estimates of the Gaussian distributions of face, non-face and lip classes in the red, green and blue (RGB) color space. The researchers then obtained visual features, namely the affine-invariant Fourier descriptors (AIFDs) [21], the DCT, the rotation-corrected DCT (rc-DCT) and the B-Spline template (BST) [19]. Due to their greater sensitivity to lip shape, the appearance-based features, DCT and rc-DCT, demonstrated good performance compared to that obtained using the shape-based features, AIFDs and BST. The system used HMM as a classifier and the performance was evaluated using the CUAVE database [19].

In [22], the authors proposed an appearance-based lip reading approach that generated dynamic visual speech features, termed the motion history image (MHI) [23], that were classified using artificial neural network (ANNs). The approach captured movement in image sequences and generated a single grayscale image to represent the whole image sequence using accumulative image subtraction techniques. However, this technique proved highly sensitive to environmental changes. In addition, information about the timing of movements was lost following the combination of sequences into a single image, resulting in lower performance. In [24], the authors reported a technique that computed the optical flow (OF) of lip motions in a video data stream. The statistical properties of the vertical OF component were used to form feature vectors suitable for training a support vector machine (SVM) classifier. However, as is the case for OF methods in general, the performance was adversely affected by sensitivity to scaling and rotation of the images.

The literature suggests that appearance-based features are generally able to produce better classification results as they carry more information, but also because of the difficulty found in shape-based approaches to extract accurate geometrical features [19]. However, the appearance-based features exhibit greater sensitivity to environmental condition changes such as illumination and head pose [14]. The main aim of the work described in this paper is to develop an approach that offers the classification performance of an appearance-based method, yet is able to provide a solution robust to environmental changes.

This paper presents a new lip reading system that has been designed to recognize speech using high-quality geometrical information obtained from the lips of the speaker. There are three novel aspects to the work. Firstly, the application of a border following technique and the construction of a convex hull was able to provide lip extraction of an accuracy that improved considerably on previously reported results. Secondly, the features extracted from the lip geometry are determined over the duration of the video sequences and pattern matching with respect to stored templates is performed according to the dynamic variations characteristic of individual word utterances. Thirdly, the lip geometry features are classified using a novel template probabilistic multi-dimension dynamic time warping (TP-MDTW) approach that is able to adapt to the differences in the way words are uttered by speakers. As far as the authors are aware, this is the first time a template probabilistic geometrical-based lip reading system has been described in the literature.

The paper is organized as follows. Section 2 briefly describes the background of lip reading systems described in the literature, Section 3 gives details of the architecture of the new lip reading system. The experimental results are given and discussed in Section 4. Section 5 provides the principal conclusions of the work and makes recommendations for future enhancements.

2. Background

There are three important issues that need to be considered when designing and implementing a lip reading system. The first is to identify the method used to locate the position of the mouth in the image; the second is the choice of visual features and the third is the speech classification method to adopt. The techniques commonly applied in research described in the literature are introduced in this section.

2.1. Mouth detection

The methods to locate the mouth can be categorized into four types: (a) template matching; (b) feature invariant; (c) machine learning and (d) knowledge based.

Template matching techniques are widely used in image processing for object detection, involving normalized cross-correlation between a known template image (an object in the training set) and an unknown new image to be classified. Such a technique has been used to detect the face [25], as well as the eyes and mouth [26].

Feature invariant methods incorporate features of target objects that do not alter even when substantial changes are made to environmental conditions such as image brightness or the pose of the subject. The methods applied include color content, object shape and motion characteristics [27].

Machine learning methods can be trained to recognize complex patterns and make intelligent decisions. Since mouth identification is often complicated by variances that arise from actions (such as opening, closing or smiling) and pose (resulting from head rotation or changes in target distance), large sample sizes are normally required to generate a high-quality machine learning data set. Popular machine learning approaches include ANNs [28], radial basis functions (RBF) [29] and the Viola Jones object recognizer [30].

Knowledge-based approaches tend to adopt techniques used by humans to locate the position of the mouth. For example, it is often reasonable to assume that the mouth is positioned in the lower part of the head below the nose and on a line of vertical symmetry. This information can be used to supplement other techniques thereby improving overall accuracy [28].

2.2. Lip visual features

Lip visual features are generally grouped into three categories [31–33]: (a) appearance-based features; (b) shape-based features and (c) a combination of both appearance and shape features.

In the extraction of appearance-based features, it is assumed that the entire mouth region is informative as far as lip reading is concerned and the region itself is isolated as a rectangle containing the mouth. As the area covered by such a mouth region can contain a large number of pixels (for example, assuming each color is represented in 8 bits, a 128×128 pixel region in RGB space will have a total of 49,152 pixels), a transformation to fewer dimensions is needed to make appearance-based approaches computationally manageable. Such transformations are typically borrowed from the image compression and pattern classification literature, such as principal components analysis (PCA) [11], the discrete cosine transform (DCT) [9], the discrete wavelet transform (DWT) [10] and linear discriminant analysis (LDA) [34].

Meaningful shape-based features for lip reading can be extracted from the inner and outer lip contours, such as the height, width, perimeter and area. These visual features can be obtained following the application of one of the parametric or statistical lip-tracking algorithms. Some popular methods for this task are edge detection [5,19], snakes [35], GVF [16], AAM [12] and the active shape model (ASM) [36].

A combination of appearance-based and shape-based features can be used in an attempt to improve the quality of the description by avoiding the main disadvantages of the two approaches, namely that appearance-based methods degrade more significantly when affected by environmental influences resulting from changes in illumination or head-pose [37] and shape-based methods generate less reliable features because of the need to extract physical measurements. For example in [33], two shape-based features (height and width) were combined with appearance-based features obtained from an intensity profile of the central mouth region expected to contain information about teeth and tongue movements.

2.3. Speech classification based on lip features

The viseme is used to describe the lip shapes, positions and movements relevant to speech in the visual domain [38]. Since only a small part of the vocal tract is visible when we speak, only partial physical information is available regarding the generation of visemes and not all can be mapped to a unique phoneme [32], the basic unit of speech in the audio domain.

A viseme may be represented by a time sequence of lip shapes, but the actual set of lip shapes and their durations are dependent on the speaker. For example, although it would be expected that the visual representation of the word 'hello' may vary between speakers (inter class), there are also likely to be differences if the word is spoken again by the same speaker (intra class), for example if on the second occasion the individual circumstances of the speaker changes, perhaps they now shout the word or they find themselves in a stressful situation.

As the application domain is the same, lip reading classification techniques are often the same as those applied in the audio speech recognition (ASR) field and, consequently, dynamic time warping (DTW) [39,40] and HMMs [10,18,41], are popular. Moreover, by using a method common to both the audio and visual aspects of speech, there is the potential for a more straightforward combination of results obtained from separate audio and visual investigations and such integration has often been carried out using machine learning techniques, such as time delay neural network (TDNN) [42], support vector machines (SVM) [43] and AdaBoost [44].

3. System architecture

The geometric-based approach for the proposed lip reading system described in this paper is shown in Fig. 1. The system can be divided into the following four stages.

- The face and then the mouth regions are extracted from the images contained in the video sequences of the speakers.
- The mouth region is segmented into lip and non-lip areas.
- A new approach is applied that uses border following and convex hull computation to extract the lip geometry and to generate shape-based features.
- A novel technique termed TP-MDTW is used to classify dynamic geometry information.

The component parts of each stage are described in detail in the following sub-sections. The software for this work was developed using Microsoft Visual C# 2010 [45] and utilized the open source image processing library, OpenCV [46].

3.1. Face and mouth detection

The speaker images as acquired from the video files are cropped to the mouth region using a face-detection process followed by a mouth-detection process. As discussed in the previous section, many techniques are available for extracting the face and mouth region. The Viola Jones object recognizer (machine learning) [30] was employed both to distinguish between the upper and lower parts of the face and to isolate the mouth region in the lower part

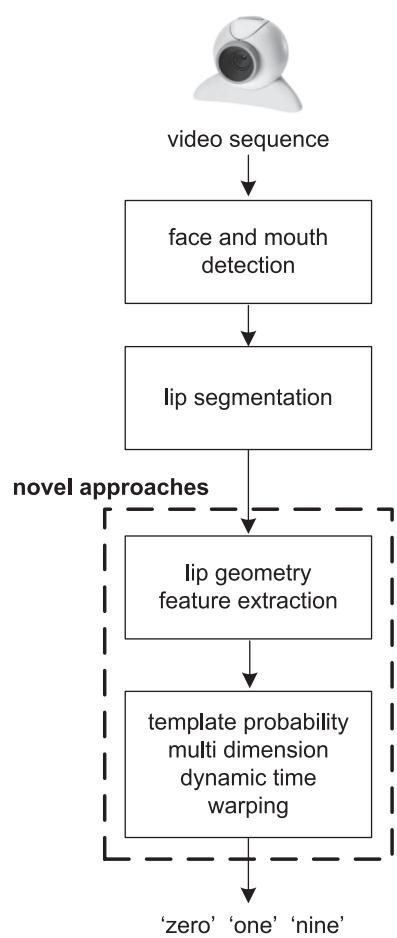


Fig. 1. Architecture of the lipreading system.

of the face, using the (knowledge-based) assumption that the mouth can be found here, as shown in Fig. 2. The Viola Jones approach was chosen because of its known robustness and a computational efficiency which allows operation in real-time on most target platforms [47].

The Viola Jones object recognizer uses simple rectangular Haar features [48] that are applied to each image in a wide range of translations and at many different scales. To select specific Haar features, Adaboost [49] is used to train a weak classifier. Single strong object classifiers can then be formed by cascading such weak classifiers as shown in Fig. 2. The advantage of weak classifiers operating in cascade is that early processing can highlight regions more likely to contain an object of interest, and these identified regions can then be subjected to concentrations of effort in subsequent operations. Also note that by adopting integral images, the Haar multiplication operations can be reduced to those involving only addition and subtraction, thereby shortening the execution time on many platforms. The approach was applied in two stages, first to obtain the face region and second to find the mouth region based on the assumption that it is located in lower half of the face as shown in Fig. 3.

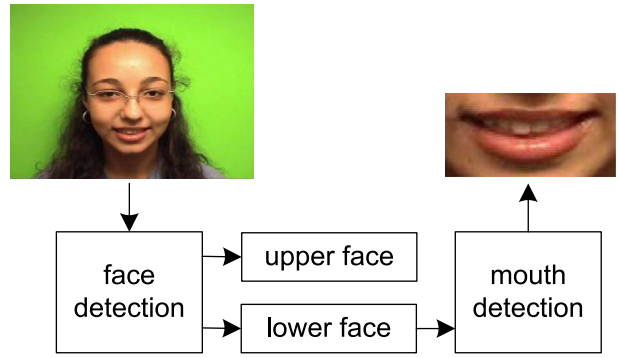


Fig. 3. Block diagram of the face and mouth detection process.

3.2. Lip segmentation

Skin detection is employed to remove the skin-colored pixels from the images and so narrow the focus to the remaining lip-colored regions. The color spaces that are in common use are red, green, blue (RGB), cyan, magenta, yellow, black (CMYK) and hue, saturation and value (HSV). Of these, the HSV color model was chosen for the segmentation operations, as this model has been found to come closest to mimicking how humans perceive color [50,51]. Furthermore, the transformation of the original RGB images to HSV is invariant to intensity at white lights, ambient light and surface orientations relative to the light source [52].

Skin colors tend to cluster in a small region of the color space. Hence, one of the easiest and most-often applied methods for lip segmentation is to define skin color cluster decision boundaries for each of the color space components. The pixel values that fall within the predefined ranges are deemed to be skin pixels and the remaining regions (including the lips) are assumed to form non-skin segments. As the lips can normally be assumed to be surrounded by skin, identification of skin pixels will isolate the mouth region. The investigations carried out in [53] found it most suitable to take as skin pixels those having hue and saturation values in the ranges $H = [0,50]$ and $S = [0.23,0.68]$ for the Far Eastern and Caucasian subjects found in the M2VTS database [54]. In the current work, the assessment of images from the CUAVE video database found that the detection of hue values alone, in particular a hue threshold of 7, was suitable for separating the skin pixels. The extracted pixels were then processed using a morphological operation (erode and dilate) to minimize high-frequency 'salt and pepper' noise content, followed by the application of a smoothing filter (down-sample and up-sample) to soften the image edges. At the end of this process a binary image was formed showing only the lip and non-lip areas as depicted in Fig. 4.

3.3. Lip geometry feature extraction

To the binary image containing the lip region a contour extraction algorithm was applied using a border following technique [55]. From the collection of contours produced, the largest is selected to generate the lip outline. Although this contour generally followed the outline of the lips, because it is generated as a simple polygon with many non-intersecting edges it remained a poor representation of the actual lip shape. However, a complex polygon such as that shown in Fig. 5(a) can be reduced to a simpler convex polygon using the convex hull algorithm [56]. This algorithm determines the convex polygon of smallest area such that it contains all the vertices of the original polygon, as shown in Fig. 5(b). Fig. 5(c) is the single final polygon. The results showed that the convex hull solution was able to extract the outline shape of the lips to an accuracy sufficient for the estimation of good quality lip geometrical features such as height, width and perimeter.

Fig. 6 shows a single video frame to illustrate the five space-based features that were obtained in the current work, namely height, width, ratio (height/width), area and perimeter. The perimeter is that of the polygon generated from the convex hull operation while the area is the region bounded by this polygon.

The geometrical information obtained from the lips was used for the lip reading process. However, the information obtained from single images is not generally able to generate good quality speech recognition results and dynamic feature information is needed to more accurately follow the time changing nature of speech. Consequently, a speaker model is needed that is also dynamic in form and which is obtained from the lip information during enrollment. If $\mathbf{G}_i$ are the geometrical features obtained from video frame i, then

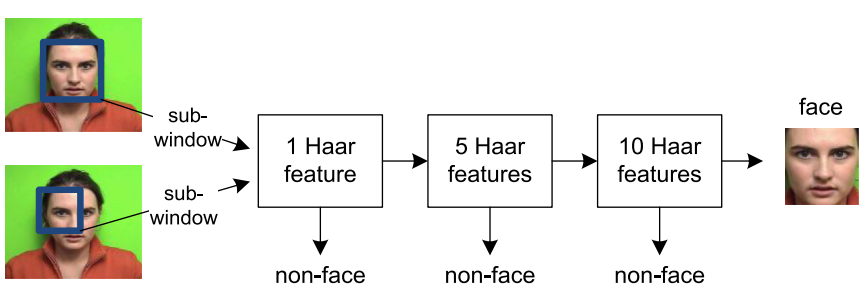


Fig. 2. Face detection using Viola–Jones object recognizer.

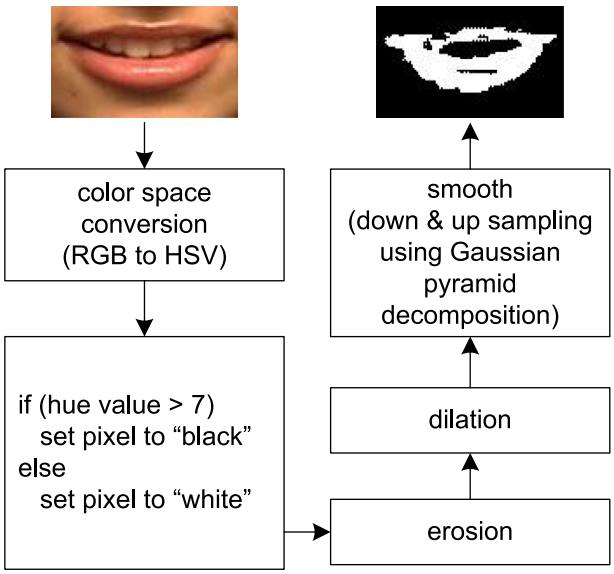


Fig. 4. Block diagram of the lip isolation process.

$$\mathbf{G}_i = [h_i \; w_i \; r_i \; a_i \; p_i]^T \tag{1}$$

where h is the height, w the width, r the ratio of height to width, a the area and p the perimeter. The geometrical features obtained from video sequence can be denoted by matrix $\mathbf{G}$.

$$\mathbf{G} = [\mathbf{G}_1 \; \mathbf{G}_2 \; \ldots] \tag{2}$$

3.4. Template probability multi dimensional dynamic time warping

The representation of the lip reading problem in the current work is now described in terms of the time series sequence shown in Eq. (2). As discussed in the previous section, DTW and HMM are the two main alternatives for time series classification and have been widely used in both speech recognition and lip reading. In addition, DTW can also be thought of as a special case of HMM, with each point along the reference signal representing a hidden state and transition probabilities restricted to prevent movements backwards in time.

DTW utilizes dynamic programming to generate candidate stretched and compressed sections in sequences of feature vectors, in order to find an alignment between two time-series that minimizes distortion [57] and in doing so produces a suitable warping function that minimizes the total distance (normally Euclidean) between an unknown sample and the reference template; While a DTW-based lip reading system has been proposed previously [39,40,58], to the best of the authors' knowledge the problem has not been addressed using multi-dimensional DTW and reference template probabilities.

The version of DTW utilized in this paper follows the approach found in [59,60]. In this approach, a two-dimensional M by N cost matrix $\mathbf{D}$ is constructed, where each of the $D(i,j)$ values is the minimum distance warped path at time i for the time series $\mathbf{x}$ and time j for time series $\mathbf{y}$, where $\mathbf{x} = (x_1,\ldots,x_i,\ldots,x_M)$ and $\mathbf{y} = (y_1,\ldots,y_j,\ldots,y_N)$ are time series. The value at $D(M,N)$ will contain the minimum distance warped path between the time series $\mathbf{x}$ and $\mathbf{y}$.

If the absolute difference between any two pairs of values in the time series $\mathbf{x}$ and $\mathbf{y}$ is given by

$$d(i,j) = |x_i - y_j| \quad 1 \leqslant i \leqslant M, 1 \leqslant j \leqslant N \tag{3}$$

then the values in the cost matrix $\mathbf{D}$ are defined as follows

$$D(1,1) = d(1,1) \tag{4}$$
$$D(1,j) = D(1,j-1) + d(1,j) \quad 2 \leqslant j \leqslant N \tag{5}$$
$$D(i,1) = D(i-1,1) + d(i,1) \quad 2 \leqslant i \leqslant M \tag{6}$$

$$D(i,j) = d(i,j) + \min \begin{Bmatrix} D(i,j-1) \\ D(i-1,j-1) \\ D(i-1,j) \end{Bmatrix} \quad 2 \leqslant i \leqslant M, \; 2 \leqslant j \leqslant N \tag{7}$$

DTW as defined in the literature is only applicable to problems requiring single feature alignment. To provide alignments in the current work that can include up to five shape-based features, the conventional approach would be to apply DTW operations to each feature separately and then subsequently select the one exhibiting the shortest distance (lowest error). However, as the features are obtained independently and there is no synchronization between them, the results obtained from the initial experiments were unsatisfactory [61].

Consequently, a further novel extension termed the multi-dimensional DTW (MDTW) was made to the DTW algorithm to allow it to operate with multiple features simultaneously in providing synchronization between the time series. The experimental results for MDTW showed a marked improvement in lip reading accuracy compared to those obtained using DTW [61]. In the MDTW method, the two time series $\mathbf{x}$ and $\mathbf{y}$ must first be reconstructed as multi-dimensional matrices where each column represents one series of the $\mathbf{G}_i$ features. These matrices can be generated from Eq. (2) and can be re-written as

$$\mathbf{x} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,M} \\ x_{2,1} & \cdots & x_{2,M} \\ \vdots & \ddots & \vdots \\ x_{K,1} & \cdots & x_{K,M} \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_{1,1} & \cdots & y_{1,N} \\ y_{2,1} & \cdots & y_{2,N} \\ \vdots & \ddots & \vdots \\ y_{K,1} & \cdots & y_{K,N} \end{bmatrix} \tag{8}$$

where M is the length of the reference video sequence, N is the length of the unknown video sequence and K is the number of geometrical features evaluated.

In this work, Eq. (3) also needs to be modified in order to operate in the multi-dimensional case and is given by

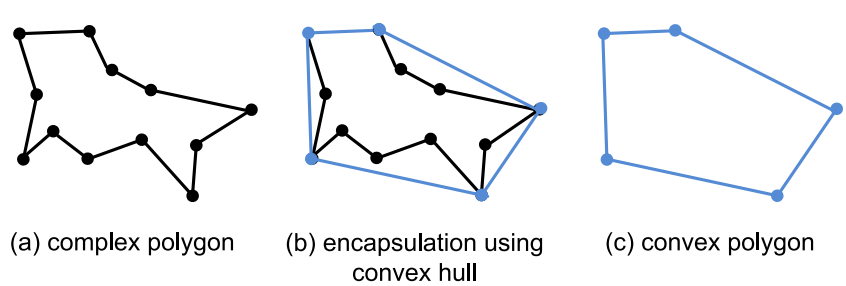


Fig. 5. Illustration of convex hull process using an 'origami cat' example.

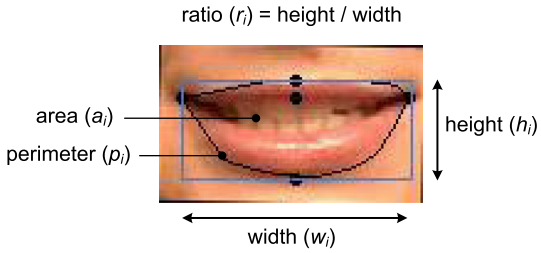


Fig. 6. Shape-based lip features obtained from a single video frame *i*.

$$d(i,j) = \sum_{k=1}^{K} \left| \frac{M \cdot x_{k,i}}{\sum_{m=1}^{M} x_{k,m}} - \frac{N \cdot y_{k,j}}{\sum_{n=1}^{N} y_{k,n}} \right| \quad 1 \leqslant i \leqslant M,\ 1 \leqslant j \leqslant N \tag{9}$$

Then the minimum distance warped path between the two multi-dimensional time series **x** and **y** using MDTW can be denoted as

$$MDTW(\mathbf{x}, \mathbf{y}) = D(M, N) \tag{10}$$

In our initial experiments, it was observed that a few of the templates used for training were never selected for the subsequent lip reading matching operations. This was found to occur because the training set contained examples of sequences spoken in a manner not found elsewhere in the training set, although they may potentially form a good match to test examples. To overcome this issue, a novel template probabilistic MDTW (TP-MDTW) technique was introduced to calculate the probability of each template being the best match to an unseen example based on the similarity between other templates in the database. The assumption is that such a template will have a greater similarity to other templates and so should be recognized as the most probable to occur and those templates having less similarity are less likely to occur.

To understand the operation of TP-MDTW, consider a system with a set of *R* reference templates $\mathbf{T_1}$, $\mathbf{T_2}$, $\mathbf{T_3}$,...,$\mathbf{T_R}$. An example is shown in Fig. 7 for the case when $R = 4$.

The similarity between the templates can be calculated using any suitable distance measurement technique. In the current implementation, MDTW is used to find the distance e_{ij} between the features in the reference templates. For any pair of unknown and reference templates $\mathbf{T_i}$ and $\mathbf{T_j}$ in the form given in Eq. (8), the distance e_{ij} between the features is found using MDTW and is given by

$$e_{ij} = MDTW(T_i, T_j) \quad 1 \leqslant i,\ j \leqslant R \tag{11}$$

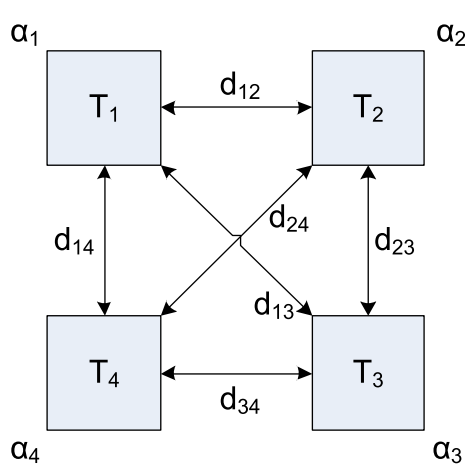


Fig. 7. Example of distance values that need to be calculated when four reference templates are defined.

Since the distances between pairs of templates is the same regardless of the starting point

$$\begin{aligned} e_{ij} &= e_{ji} \quad j \neq i \\ e_{ij} &= 0 \quad j = i \end{aligned} \tag{12}$$

To find the similarity between any one given template and the remainder, the cumulative distance to other templates can be computed by

$$\alpha_i = \sum_{j=1}^{R} e_{ij} \quad 1 \leqslant i \leqslant R \tag{13}$$

Based on the cumulative distance between the templates, the probability of a template being the best match to an unseen example can be calculated using

$$P(T_i) = \frac{\frac{1}{\alpha_i}}{\sum_{g=1}^{R} \left[\frac{1}{\alpha_g}\right]} \quad 1 \leqslant i \leqslant R \tag{14}$$

where

$$\sum_{i=1}^{R} P(T_i) = 1 \tag{15}$$

To make best use of the alternative approaches that are available at this stage, these have been implemented as models able to process inputs (template probability, reference template and unknown sample), as shown in Figs. 8 and 9. This work introduces four lip reading models whose operations were designed based on a series of preliminary experiments.

The objective of model 1 is to determine the accuracy of the system assuming there is a uniform probability in the distribution for each template as shown in Eq. (16).

$$P_1(T_i) = \frac{1}{R} \quad 1 \leqslant i \leqslant R \tag{16}$$

The output for model 1 for an unknown sample ϕ is

$$O_1(\varphi) = \sum_{i=1}^{R} [P_1(T_i) \cdot MDTW(\varphi, T_i)] \tag{17}$$

Model 2 selects the template with the largest probability of the all reference templates in the database by using Eq. (18).

$$P_2(T_i) = \begin{cases} 1 & \text{if } P(T_i) = \max\limits_{1 \leqslant k \leqslant R} \{P(T_k)\} \\ 0 & \text{otherwise} \end{cases} \quad 1 \leqslant i \leqslant R \tag{18}$$

The assumption is that this template gives the closest representation to the features from the unknown example. The output for model 2 is given by

$$O_2(\varphi) = \sum_{i=1}^{R} [P_2(T_i) \cdot MDTW(\varphi, T_i)] \tag{19}$$

The objective of model 3 is to measure performance based on the template probability that is calculated from the accumulated distance from other templates as found by Eq. (14). The output of model 3 is given by

$$O_3(\phi) = \sum_{i=1}^{R} [P(T_i) \cdot MDTW(\phi, T_i)] \tag{20}$$

Model 4 measures performance based on the template probability calculated from the accumulated distance, but also gives emphasis to the reference template that has the highest probability. Its performance calculation is given by

$$O_4(\phi) = \sum_{i=1}^{R} [(P(T_i) + P_2(T_i)) \cdot MDTW(\phi, T_i)] \tag{21}$$

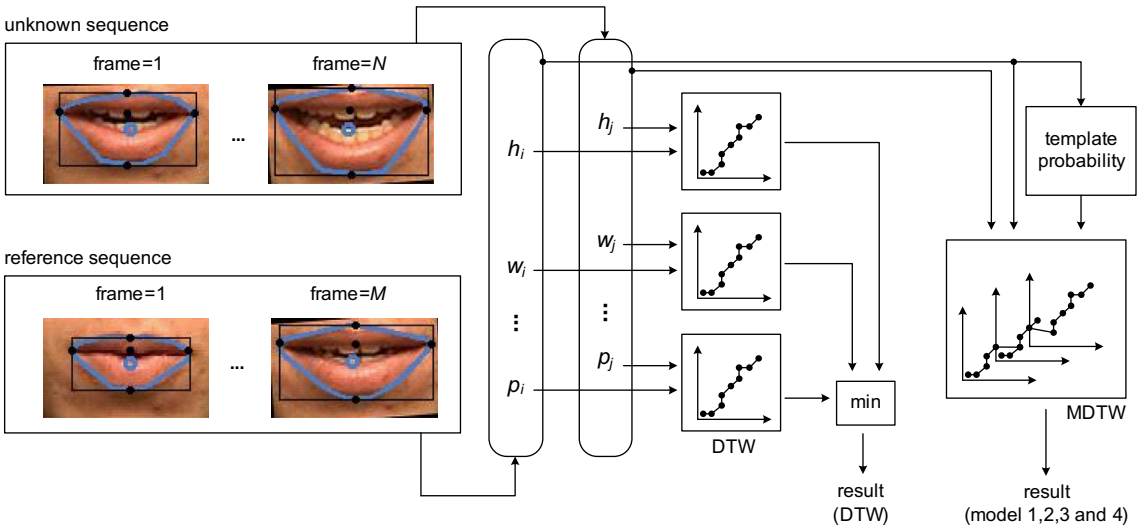


Fig. 8. General structure of the classification operations used in the lip reading system.

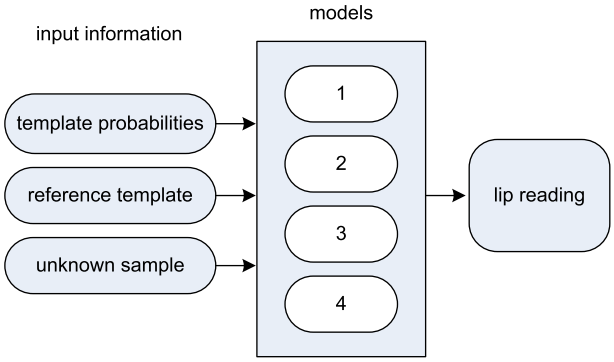


Fig. 9. Models used in the lip reading classification.

Clearly, the number of models is not limited to these four cases and further models could be developed to implement additional operations.

4. Results and discussion

In order to investigate the effectiveness of the proposed approaches in practical applications, the CUAVE corpus database was used to provide examples of speech and video sequences of the speakers [19]. The database consists of 7000 utterances of connected and isolated digits spoken by 36 individuals, where 19 speakers are male and the remainder are female. The speakers also have a range of skin and lip tones as well as face and lip shapes and a number of the subjects wear additional visual features such as glasses, facial hair, and hats. Lighting was controlled and a green background was employed to allow custom video backgrounds to be added using chroma-keying if required. The video sequences were recorded at a resolution of 720 × 480 in MPEG-2 format at 29.97 frame/s and encoded at a data rate of 5000 kbit/s.

4.1. Face and mouth detection

Fig. 10 shows sample results of the face and mouth detection process. Fig. 10(a) shows initial face detection using the original database image and Fig. 10(b) shows Voila Jones method that used to identify candidate mouth regions. By incorporating the knowledge that the mouth can be found in the lower half of the face, as shown in Fig. 10(c), the calculation time needed to isolate the mouth region is effectively halved and also reduces the risk of false detection that can arise from the erroneous classification of the eyes as a mouth. The resulting detection of the mouth is shown Fig. 10(d). Fig. 10(e) shows an image that has been superimposed using chroma-keying to replace the green background with one more complex and, although there are three faces visible in the image, the algorithm selects the largest face region in the frame based on the assumption that the target speaker is likely to be the one nearest to the camera. Fig. 10(f) shows the detection process is able to perform correctly for a speaker wearing a hat.

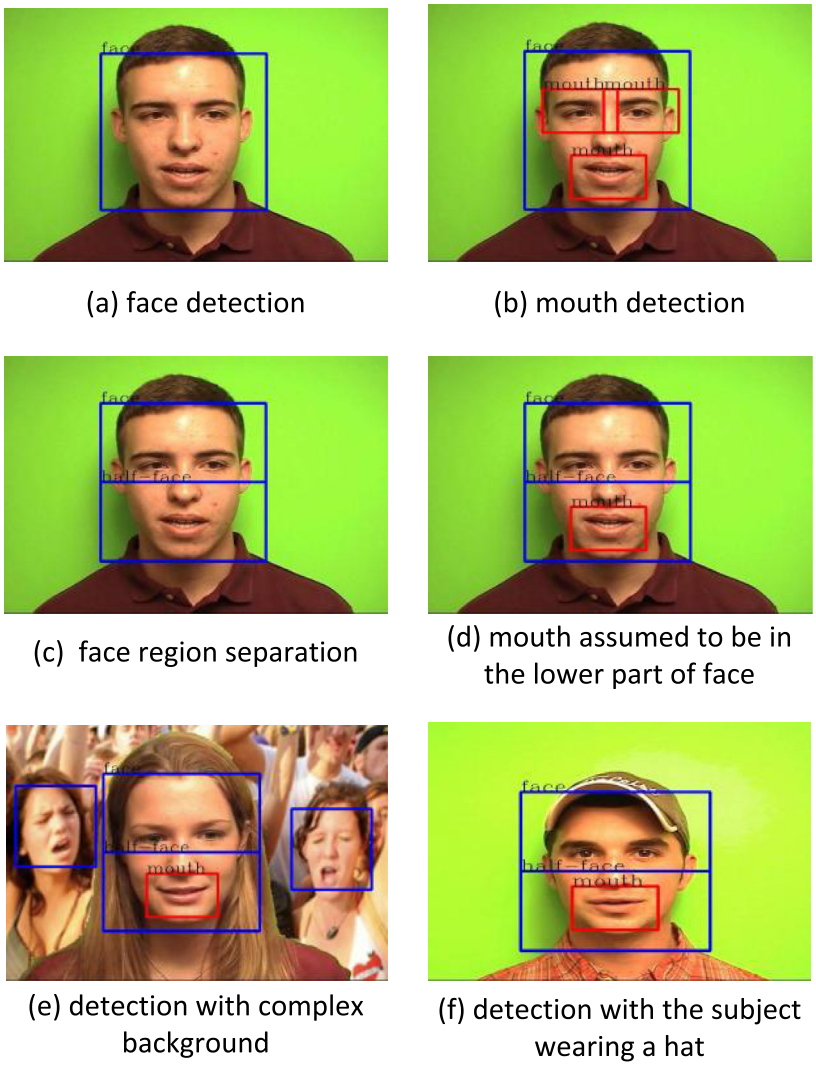
4.2. Lip segmentation

Values of hue suitable for lip segmentation were determined using a two-stage qualitative evaluation approach. In the first stage, it was found that hue values outside the range 6–9 consistently yielded poor segmentation results. In the second stage, hue values within this range were investigated in further experiments carried out on a data set containing 180 face images from the CUAVE database (five for each speaker) selected to exhibit a range of facial expressions

For the qualitative assessment a visual inspection is normally conducted; for example in [62] four benchmark points were defined for the lips (left, right, top and bottom) and their locations were used to grade performance as 'wrong', 'poor', 'fair', 'good' or 'perfect'. In the current work, the four grades used to define performance are 'wrong', 'poor', 'satisfactory' and 'good', with the selection depending on how closely the lip region boundary has been determined and the accuracy of fit of the four benchmark points (left, right, top and bottom of the lips). Fig. 11 shows examples of grading and classification carried according to the scheme shown in Table 1. It was found that the detection of hue values alone, where the threshold was set at a hue value of 7, was appropriate to separate the skin pixels in mouth region. The number of images that resulted in good visual lip segmentation quality is over 75% with no image wrongly segmented (as the mouth region was found in all cases).

4.3. Lip geometry feature extraction

Fig. 12 shows the results of lip geometry extraction from color images for three different speakers. Binary lip images, as shown



(a) face detection

(b) mouth detection

(c) face region separation

(d) mouth assumed to be in the lower part of face

(e) detection with complex background

(f) detection with the subject wearing a hat

Fig. 10. Face and mouth detection process.

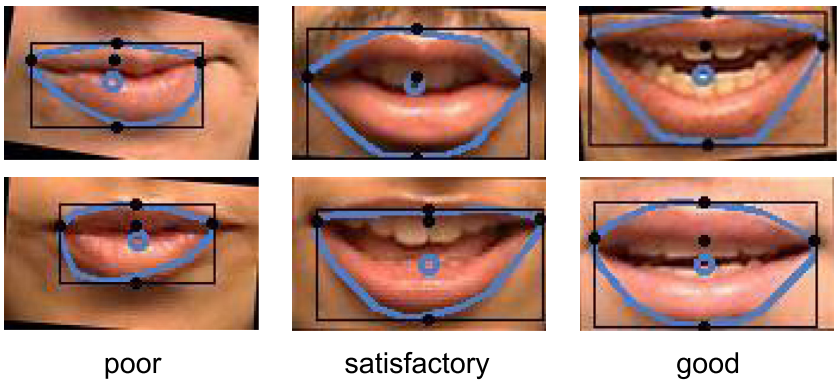


Fig. 11. Examples of the grading classification used in the qualitative assessment.

in Fig. 12(b), are generated using the convex hull process shown in Fig. 5. A border following technique [55] was used to generate a collection of contours with different size and shape throughout the border of binary image as shown in Fig. 12(c). Our experiments show that the largest contour generated by this process will most closely contain the lip area as shown in Fig. 12(d). Using the largest contour, it can be seen that the convex hull approach is able to generate a close approximation to the actual lip shape in the original image.

In order to obtain good quality height and width information for later classification purposes, the lip shape must be consistently aligned. As the image used in this paper involved only frontal pose, alignment can be achieved by rotation so that the left and right vertices of the lip contour are at the same vertical position in the

Table 1
Qualitative evaluation of lip classification.

Hue value	Grade classification (%)			
	Good	Satisfactory	Poor	Wrong
>6	43.9	26.7	29.4	0
>7	75.0	18.9	6.1	0
>8	50.0	25.0	25.0	0
>9	21.1	16.7	62.2	0

image, as shown in Fig. 13. The vertical and horizontal dimensions of the bounding box that encapsulates the entire shape represents the lip height and width respectively while the area and perimeter can be extracted from the convex polygon.

To produce an initial confirmation of the performance of the new approach, it was compared to both the GVF technique previously used to extract geometrical-based information from the lip region in [14] and the active contour ('snakes') approach as found in [63]. The binary lip image in Fig. 12(b) for speaker 's04f' using GVF and active contour techniques can be seen in Fig. 14 and 15. For this example, both techniques were unable to convergence to the lip shape, despite repeating the experiments using a range of different configuration parameters. The main problems encountered were that the pattern of external energy generated was not closely related to the shape of the mouth and the outcome was unreliable, depending greatly on the initialization region and the termination criteria set during detection. In addition, the calculation time taken to converge to a final result consistently took longer than the convex hull method and the actual calculation time needed varied significantly and unpredictably depending on

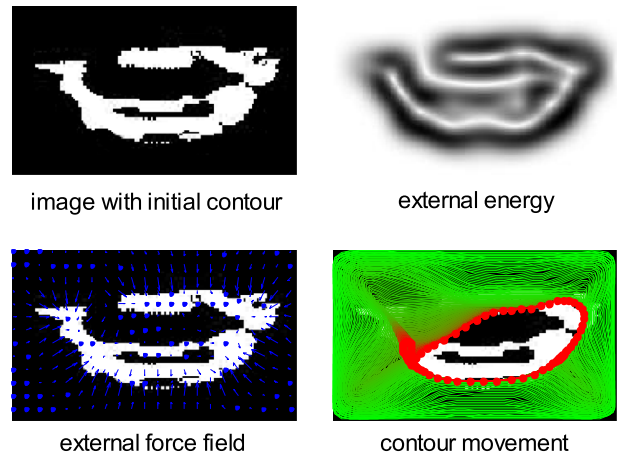


Fig. 14. Lip contour detection using the Gradient Vector Field (GVF) method for speaker 's04f'.

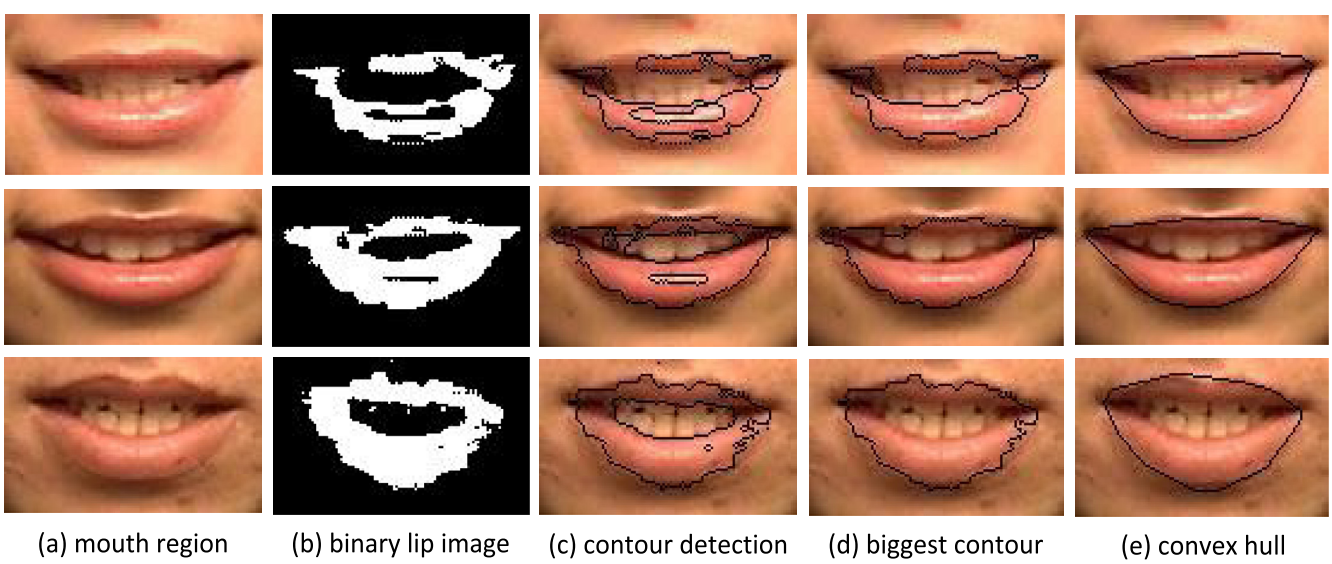


Fig. 12. Convex hull results for subject 's04f' (top row), 's05f' (middle row) and 's01 m' (bottom row).

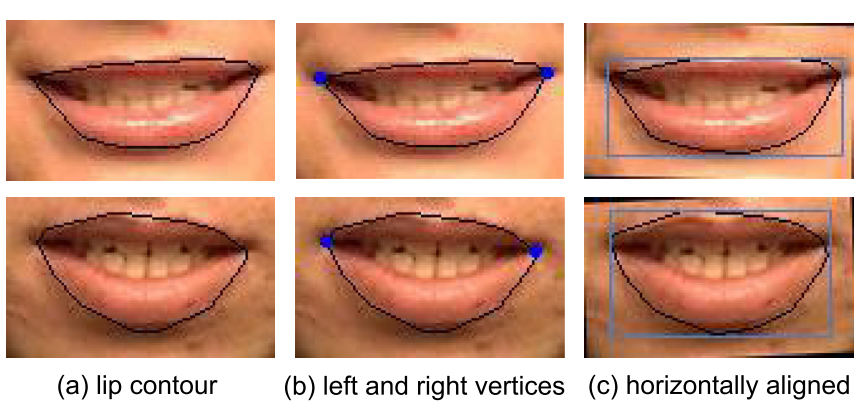


Fig. 13. Automatic alignment using left and right vertices.

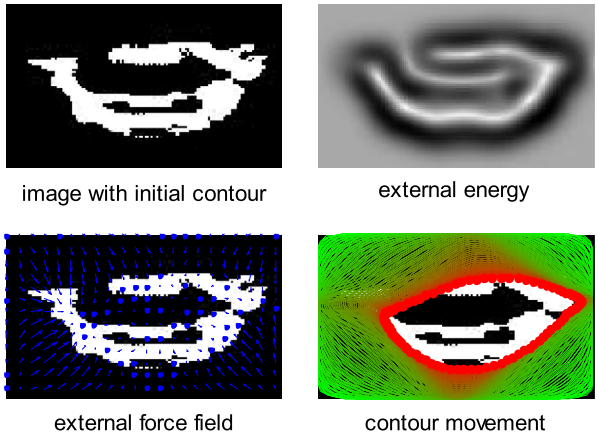


Fig. 15. Lip contour detection using Active Contours ('snakes') method for speaker 's04f'.

the parameters chosen. The results present here for the convex hull method also visually performed better than BST technique described in [19] when using the same data corpus.

Quantitative evaluations were conducted to assess the accuracy of the lip extraction method used in this paper to generate geometrical features. For this evaluation, a data set containing five different facial expressions for each speaker (a total of 180 images) was established from the 36 speakers in the CUAVE database.

The method used for the quantitative evaluation was based on that described in [64], where four key lip points (top, bottom, left and right) are defined in both the original image and the convex hull contour. The distance between corresponding points (in pixel units) defines an error that is normalized according to the distance between the mouth corners. For comparison purposes, the ASM technique proposed by Cootes et al. [36] was also used to determine the same four lip points in the original image. The ASM technique has a tendency not to be able to escape from local minima and to produce unreliable results under certain initial conditions. So, for fair comparison, the shape model was pre-initialized using prior information regarding the location of the eyes and mouth obtained from the Viola–Jones object recognizer used in our system. Fig. 16 shows the lip outline recognition performance for the convex hull and ASM methods when compared with a manual annotation of the lip outlines. In Fig. 16(a) and (b), the shape produce by the convex hull matches the outline of the lips, while the ASM results show errors compared to the expected shape both for the top and the bottom lip. Fig. 16(c) shows an example in which the two techniques exhibit similar performance.

Table 2 shows a quantitative evaluation of the convex hull and ASM techniques for 180 image samples. It shows that, compared to a manual annotation of the lips, the convex hull method has an error of 7.5% in height and 3.2% in width. The fact that the height error is larger, is likely to be due to a small number of the subject images in the database having a shadow under their bottom lip because of the illumination angle (an example of this can be seen for the first 'satisfactory' speaker in Fig. 11), thus increasing the uncertainty in the determination of the height measurement. Nevertheless, the quantitative results also confirm that the lip extraction approach based convex hull is suitable for determining features. The results shown in Table 2, demonstrate that the ASM approach is significantly worse than the Convex Hull method at extracting the outline of the lips.

The geometrical information obtained from the lips is used as input to the lip reading process, but the performance can be substantially improved using information not just from single images, but from dynamic information generated from a sequence of feature values obtained during the speech utterance. By finding the convex hull contour for each video frame, a time-series of feature values can be derived, as shown in Fig. 17. This information is stored as speaker models (templates) during training for later use in the lip reading system in which the models are compared with test series.

Fig. 18 shows the time-domain changes in the height, width and ratio of the lips when a number of different speakers uttered the

Table 2
Quantitative evaluation of the lip classification.

Method	Relative errors (%)	
	Height	Width
Convex hull	14.42	6.85
ASM	21.95	10.09

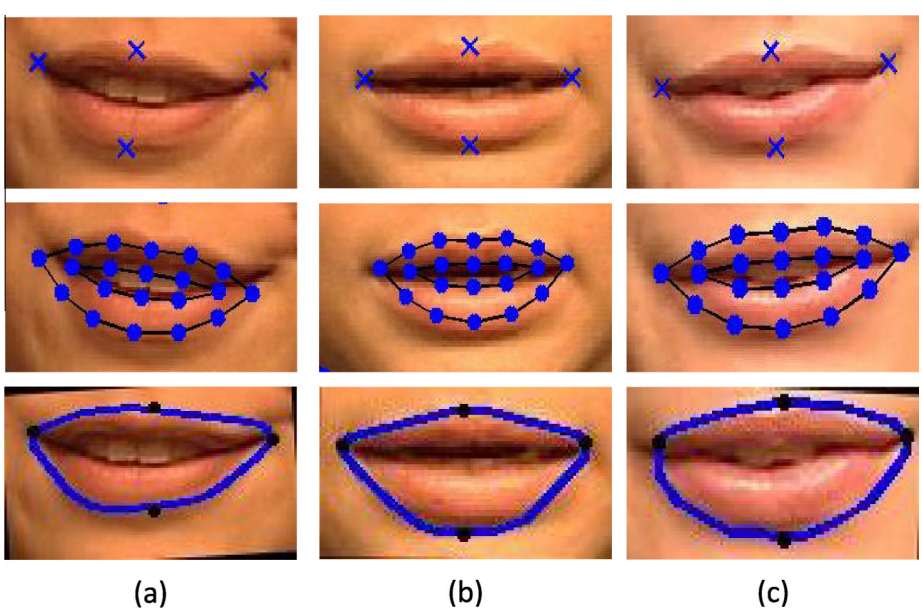


Fig. 16. Comparison between manual annotation (top row), ASM technique (center row) and convex hull technique (bottom row) for subject (a) 's28f', (b) 's34f' and (c) 's04f'.

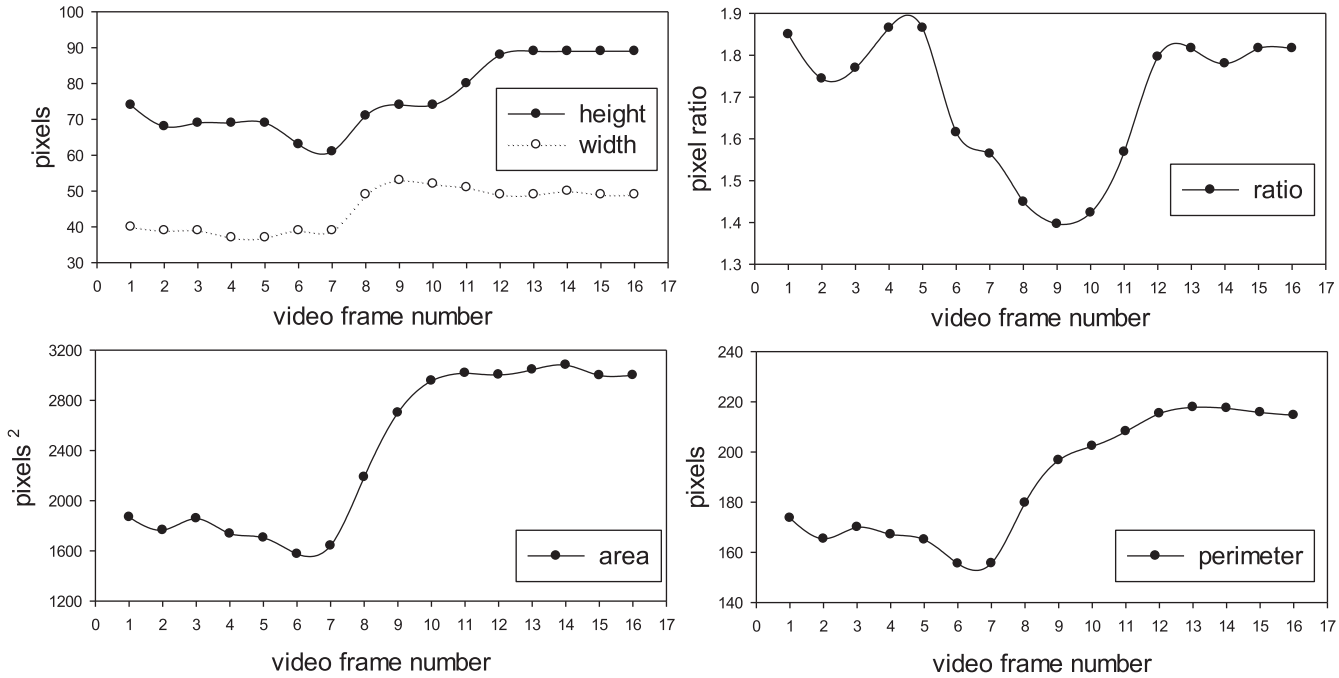


Fig. 17. Dynamic lip information for digit 'one' uttered by speaker 's01 m' in the CUAVE database.

digit 'five'. The results confirm that a similar underlying time-varying pattern was produced by all the speakers, indicating the potential of this method in its ability to identify the words being spoken.

Calculation time is also an important consideration in lip reading systems, particularly for real-time recognition in which the elapsed processing time must be less than the time interval between consecutive frames of the video. The CUAVE database has an interval between frames of 33.37 ms and Fig. 19 demonstrates that our proposed system is able to complete its operations within this time when running 64-bit Windows 7 on a 3.2 GHz Intel i5 processor and 4 GB memory. The processing time for producing the results for each of the subjects is comfortably within the frame interval time and shows good repeatability.

4.4. Classification

The CUAVE database consists of five sessions, in each of which the subject speaks the words 'zero' to 'nine'. In the investigations, data from sessions 1, 2 and 3 (30 samples) were employed for training and the data from sessions 4 and 5 (20 samples) were used for testing. All the 36 speakers from the database were used in this work making a total of 1800 samples for use in demonstrating the utility of the new approach. Two investigations have been designed to measure the performance of the lip reading system, namely classification using single features and classification using multiple features.

The classification performance using the TP-MDTW technique for the five lip geometrical features operating individually, namely height, width, ratio of height to width, area and perimeter, are shown in Table 3. Model 1 provides the ground truth regarding the accuracy of the system, employing a uniform probability distribution for each template. Of the models investigated, it can be seen that model 4 (based on the template probability calculated from accumulative distance) provides the best single feature classification. It can be seen that using lip area feature produced the best performance, providing almost 62% correct lip reading identification using model 4.

Investigations of the classification results obtained using various combinations of lip geometry features were carried out to improve further the lip reading system performance. The results of the classification using TP-MDTW shown in Table 4 demonstrate that the combination of height, width and ratio information gave the best performance, improving the classification performance up to 70.69% correct when using model 4.

To demonstrate the performance of TP-MDTW, a comparison with existing DTW and HMM classifiers was made using model 4. Left–right HMM models with eight states were used to develop word models, each state having an observation probability distribution modeled by a single Gaussian with diagonal covariance. The same lip images used to extract lip geometry for TP-MDTW were used to generate the DTW and HMM results and the recognition results are shown in Fig. 20. Compared to DTW and HMM, the classification results show a significant improvement, and the combination of height, width and ratio (HWR) performed the best, with 70.69% of the classification being successful using model 4, and the corresponding figures being 55.97% for HMM and 52.22% for DTW. From the results, it can be seen that the simple measures of height and width and their ratio was sufficient to represent the lip dynamic information and proved suitable for the lip reading system.

In order to analyze the recognition results, hypothesis testing based on McNemar's test [65] was used to decide whether the differences in performance between two algorithms applied to the same database is statistically significant. By using McNemar's test, it was found that the performance difference between TP-MTDW using Model 4 and the baseline DTW classifier is significant at the 0.001 level. Moreover, a comparison between TP-MTDW using Model 4 and the HMM classifier also exhibited a significance in performance difference at a level of 0.001.

To assess the performance of the new shape-based approach with respect to motion-based and appearance-based techniques, both OF and DCT methods were implemented in the manner described in [24] and [66]. The same lip images used to extract the lip geometry were supplied to the OF and DCT implementations, producing respectively recognition rates of 26.94% and

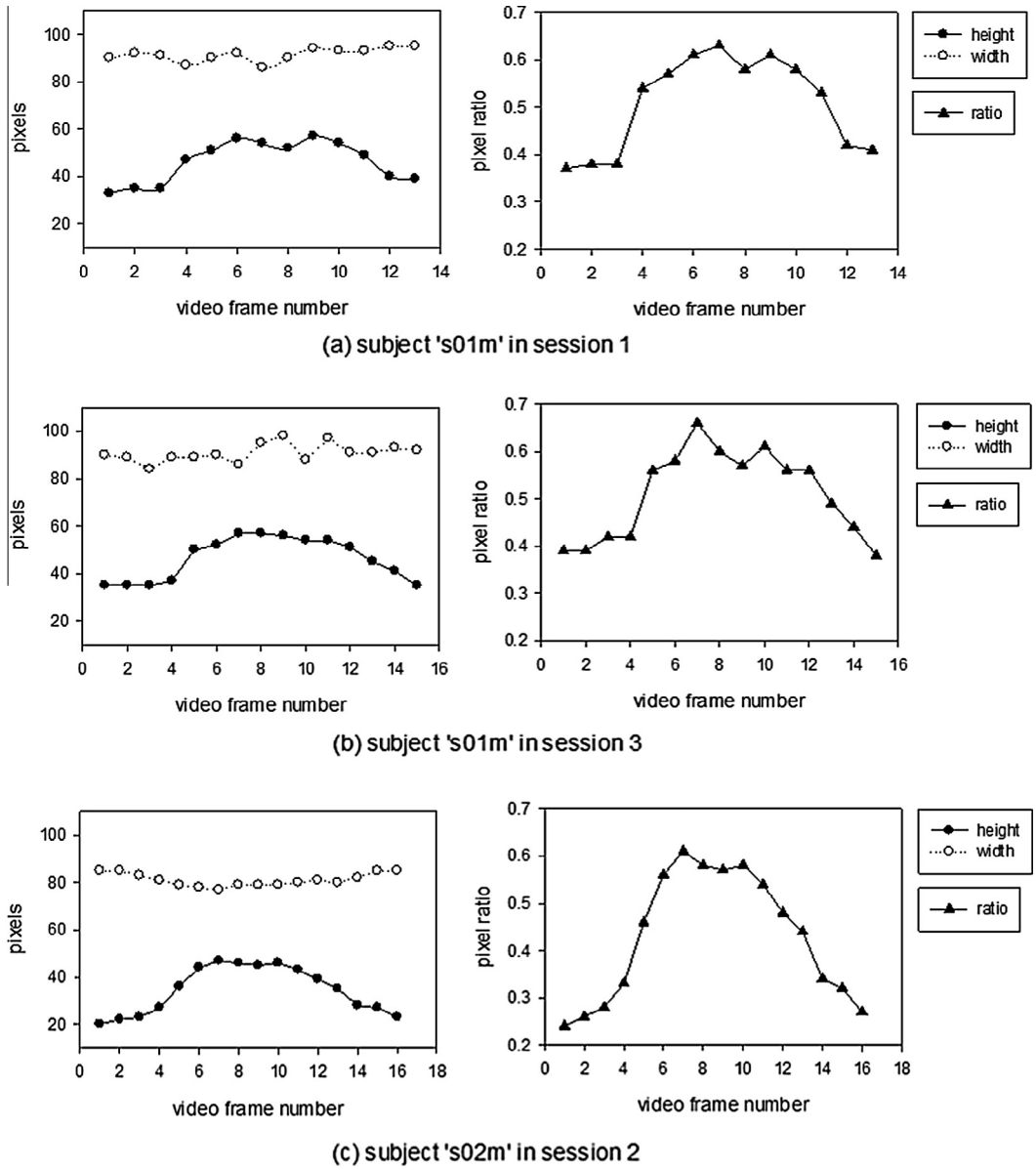


Fig. 18. Dynamic lip information showing changes in lip height, lip width and the ratio of height to width for digit 'five' obtained from the CUAVE database.

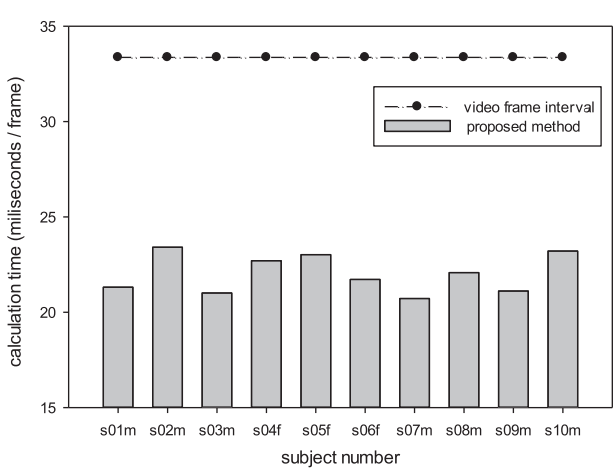


Fig. 19. Processing time using the proposed method for the first 10 subjects in the CUAVE database.

Table 3
Word accuracy for single lip geometry features using the TP-MDTW approach.

Model type	Geometry features				
	Height (%)	Width (%)	Ratio (%)	Area (%)	Perimeter (%)
Model 1	53.33	40.28	44.17	57.36	48.75
Model 2	53.75	39.44	44.44	59.58	52.64
Model 3	54.31	41.94	45.00	59.72	52.64
Model 4	56.53	42.78	48.06	61.81	55.83

57.92%. Using McNemar's test, the shape-based results exhibited a significant difference from the OF and DCT implementations, both at the 0.001 level. Fig. 21 shows a direct comparison of the confusion matrices from digit '0' until '9' obtained for OF, DCT and HWR features classified using TP-MDTW (model 4). It is important to note that the recognition results obtained in this work only use three geometrical features (height, width and ratio), while OF and DCT require 8192 features and 16 features respectively.

Table 4
Word accuracy for candidate combinations of lip geometry features classified using TP-MDTW.

Model type	Geometry feature combinations[a]					
	HW (%)	HWR (%)	AP (%)	HWRA (%)	HWRP (%)	HWRAP (%)
Model 1	55.56	61.39	57.36	57.78	54.58	58.19
Model 2	56.81	62.78	59.44	60.83	57.64	61.11
Model 3	57.50	65.42	59.72	59.44	56.67	59.86
Model 4	60.42	70.69	61.94	62.64	58.75	63.33

[a] H = height, W = width, R = ratio, A = area, P = perimeter.

For any classification approach, an important issue is that of scalability. It is known that extra calculation time will be needed to accommodate applications that may require additional features or dimensions and it is important to demonstrate that TP-MDTW is capable of such an extension. We used the determination of OF motion-based features to estimate the effect of scalability, with the number of features being increased in stages to 200 and each timing measurement carried out 10 times. Fig. 22 shows the time needed for pair-wise comparison of the features in TP-MDTW and it can be seen that the time taken to complete the underlying operations is only 0.442 ms, with an additional 3.477 μs needed for each additional feature. It is clear that the calculation time increases approximately linearly with the number of features. For the application of TP-MDTW to the lip reading application reported in this paper, classification involves just three shape-based features and it took 0.45 ms to complete the operation. In comparison, the OF motion-based method with 8192 features took approximately 29 ms, while the DCT appearance-based with 16 features took around 0.49 ms. As appearance-based and motion–based methods generally use far more features than shape-based methods, the latter approach is generally much less computationally intensive.

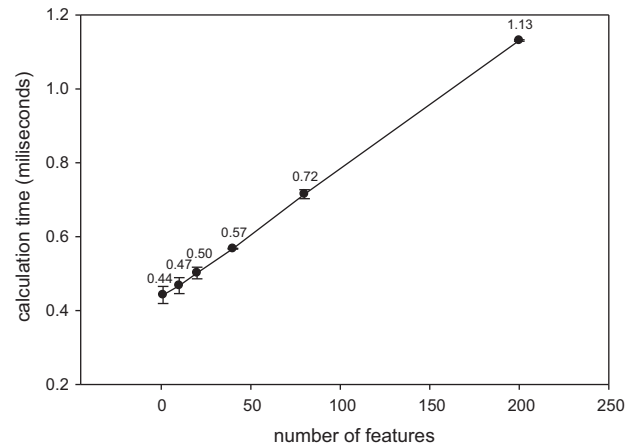


Fig. 22. Mean calculation times of pairwise comparisons of the features in TP-MDTW using samples from speaker 's01 m' in session 1. The error bars indicate ±1 standard deviation for measurements obtained using 1, 10, 20, 40, 80 and 200 features.

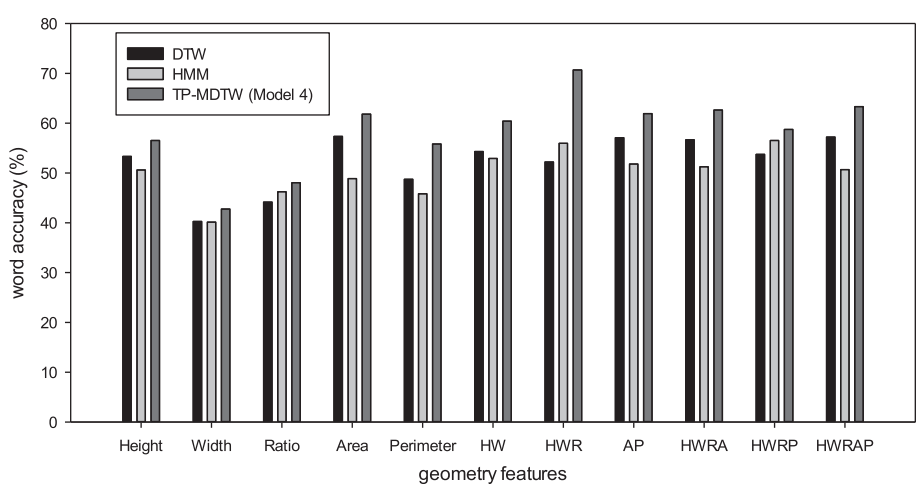


Fig. 20. Performance of different classifiers using single and combinations of geometrical features.

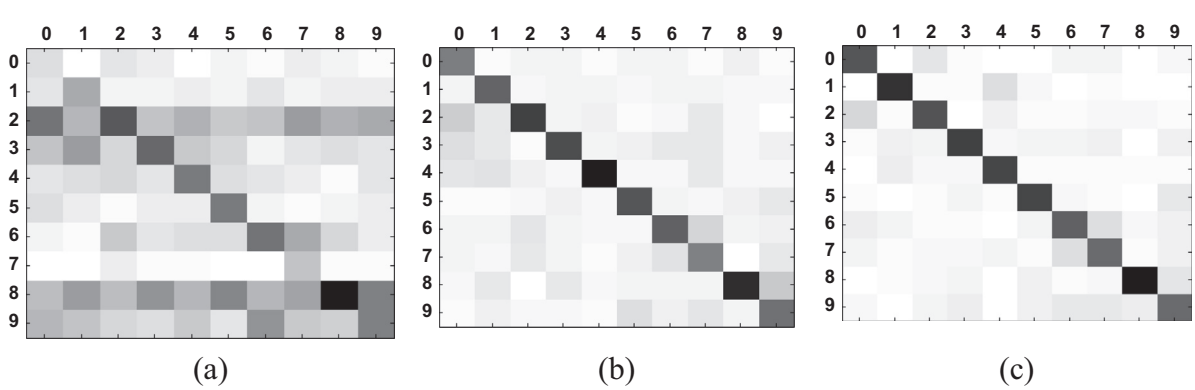


Fig. 21. Confusion matrices using (a) OF, (b) DCT and (c) HWR features.

4.5. Comparison with other studies

The studies carried out by Benhaim et al. [67] using the CUAVE database reported a speech recognition accuracy of 85% in speaker independent experiments. For visual features the approach used histogram-based descriptors around twelve lip landmarks determined using an AAM fitting technique and the classification involved multiple kernel learning and SVM. Similar results were reported by Papandreou et al. [68] who achieved a best recognition rate of 83% in speaker independent experiments when using AAM visual features obtained from the entire lower face with six shape and six texture coefficients and when using HMM for classification.

Both performance figures are better than those obtained in the current study, and this can largely be attributed to the greater number of visual speech parameters used in the approaches, including non-lip feature points, histogram-based information, as well as differences in the visual classification method. Both works utilized an AAM fitting technique to extract visual speech features, including those in the mouth region. Although AAM is known to perform well in extracting visual face features, the shape and appearance of the objects of interest are learned from a training set that requires the manual annotation of a very large number of images in order to construct a suitable statistical model. Such a supervised learning approach is extremely tedious and time consuming, and is in contrast to the approach presented in our work that implements an unsupervised learning method to accomplish a robust and accurate segmentation of the lips. Furthermore, the approach introduced in this paper uses a significantly less computationally intensive method allowing the classification to be achieved in real time.

Furthermore, the performance of the AAM approach is known to suffer if the target object is partly occluded, as the fitting algorithm has the tendency to become stuck in local minima. This problem is likely to be particularly apparent if the set of target objects has a large variability as will occur for human faces, not only due to the wide range of facial shapes and features in a normal population, but also because subjects are known to wear spectacles, make-up or beards, whose variability cannot normally be fully captured at the training stage. In the method presented in this paper, the region of interest (lip region) is identified using a Viola–Jones object detector whose application is substantially independent of image variability at the test stage.

5. Conclusion and future work

This paper has described a new approach for extracting lip geometry from the mouth region using a skin color filter followed by a border following method and the application of a convex hull technique. Compared to earlier techniques, this solution has the advantage of reliably extracting the lip contour without needing to make any a priori geometrical model assumptions while still being able to processes images in real time. Five geometrical features were extracted from each image in CUAVE database which are height, width, ratio, area and perimeter and these were used as input to TP-MDTW, a novel technique to classify lip dynamic geometry information using the probability of matching a reference template in the database while performing multi-dimensional DTW. Four models have been proposed to work with this technique.

The goal stated in the introduction, namely of delivering a shape-based model with the performance of appearance-based method has been achieved. This has been demonstrated in the performance of the new method in the lip reading of 10 English digits available in the CUAVE database. The experiments showed that the proposed method provides a high performance lip-feature extraction technique and that TP-MDTW is a promising classifier for lip reading.

Although performance of the lip reading system presented in this work achieved best recognition results just using three lip geometry features (height, width and its ratio), the potential exists to further enhance the current system by including additional visual parameters especially in lower face to augment those already presented, with the aim of improving the performance and robustness of the lip reading system as well as providing a high-performing visual component in an audio-visual speech recognition system.

As a future work, we are currently investigating scale invariant features based on the lip geometry. One possible scale invariant feature is the multiple lip angles that can be derives directly from the lip height and width. A further advantage of using lip invariant features is that speaker independent experiments can be performed more easily as the features are effectively normalized automatically for all the speakers in the database. We are also investigating the integration of visual feature into speech recognition system and analyzing its performance under noisy condition.

References

[1] J. Aron, How innovative is Apple's new voice assistant, Siri?, The New Sci 212 (2836) (2011) 24.
[2] W. Kim, J.H.L. Hansen, Feature compensation employing variational model composition for robust speech recognition in in-vehicle environment, in: J.H.L. Hansen, P. Boyraz, K. Takeda, H. Abut (Eds.), Digital Signal Processing for In-Vehicle Systems and Safety, Springer, US, 2012, pp. 175–185.
[3] H. McGurk, J. MacDonald, Hearing lips and seeing voices, Nature 264 (1976) 746–748.
[4] G. Potamianos, C. Neti, Improved ROI and within frame discriminant features for lipreading, in: International Conference on Image Processing, vol. III, 2001, pp. 250–253.
[5] X. Zhang, C.C. Broun, R.M. Mersereau, M.a. Clements, Automatic speechreading with applications to human–computer interfaces, EURASIP J. Adv. Signal Process. 2002 (11) (2002) 1228–1247.
[6] E. Benhaim, H. Sahbi, G. Vitte, Designing Relevant features for visual speech recognition, in: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013.
[7] Q. Summerfield, Use of visual information for phonetic perception, Phonetica 36 (4–5) (1979) 314–331.
[8] G. Potamianos, C. Neti, J. Luettin, I. Matthews, Audio-visual automatic speech recognition: An overview, in: Issues in Visual and Audio-Visual Speech Processing, MIT Press, 2004.
[9] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, Large-vocabulary audio-visual speech recognition: a summary of the Johns Hopkins Summer 2000 Workshop, in: Proc. Works. Multimedia Signal Processing, 2001, pp. 619–624.
[10] G. Potamianos, H.P. Graf, E. Cosatto, An image transform approach for HMM based automatic lipreading, in: International Conference on Image Processing. ICIP98, vol. 3, 1998, pp. 173–177.
[11] S. Dupont, J. Luettin, Audio-visual speech modeling for continuous speech recognition, IEEE Trans. Multimedia 2 (3) (2000) 141–151.
[12] T.F. Cootes, G.J. Edwards, C.J. Taylor, Active appearance models, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 681–685.
[13] L. Rabiner, B.-H. Juang, Fundamentals of Speech Processing, Prentice Hall Signal Processing Series, 1993.
[14] P.S. Aleksic, G. Potamianos, A.K. Katsaggelos, Exploiting visual information in automatic speech processing, in: Handbook of Image and Video Processing, 2005.
[15] I.S. Pandzic, R. Forchheimer, MPEG-4 Facial Animation: The Standard, Implementation and Applications, vol. 13, John Wiley and Sons, 2002. no. 5.
[16] C. Xu, J.L. Prince, Gradient vector flow: a new external force for snakes, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1997, pp. 66–71.
[17] J. Canny, A computational approach to edge detection, IEEE Trans. Pattern Anal. Mach. Intell. 8 (6) (1986) 679–698.
[18] T. Chen, Audiovisual speech processing, IEEE Signal Processing Magazine, vol. 18, IEEE, 2001, pp. 9–21. no. 1.
[19] E.K. Patterson, S. Gurbuz, Z. Tufekci, J.N. Gowdy, Moving-talker, speaker-independent feature study, and baseline results using the CUAVE multimodal speech corpus, EURASIP J. Adv. Signal Process. 2002 (11) (2002) 1189–1201.
[20] S. Theodoridis, K. Koutroumbas, Pattern Recognition, second ed., vol. 8, Academic Press, 2003. no. 3.
[21] S. Gurbuz, Z. Tufekci, E. Patterson, J.N. Gowdy, Application of affine-invariant Fourier descriptors to lipreading for audio-visual speech recognition, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, 2001, pp. 177–180.

[22] W. Yau, D. Kumar, S. Arjunan, Voiceless speech recognition using dynamic visual speech features, in: HCSNet Workshop on the Use of Vision in HCI, 2006.
[23] A.F. Bobick, J.W. Davis, The recognition of human movement using temporal templates, IEEE Trans. Pattern Anal. Mach. Intell. 23 (3) (2001) 257–267.
[24] A.A. Shaikh, D.K. Kumar, W.C. Yau, M.Z.C. Azemin, J. Gubbi, Lip reading using optical flow and support vector machines, in: 3rd International Congress on Image and Signal Processing (CISP), 2010, pp. 327–330.
[25] Z. Jin, Z. Lou, J. Yang, Q. Sun, Face detection using template matching and skin-color information, Neurocomputing 70 (4–6) (2007) 794–800.
[26] J. Rurainsky, P. Eisert, Template-based eye and mouth detection for 3D video conferencing, in: Proceedings of the International Workshop on Very Low Bitrate Video, 2003, pp. 23–31.
[27] Y.-L. Tian, T. Kanade, J. Cohn, Robust lip tracking by combining shape, color and motion, in: Proceedings of the 4th Asian Conference on Computer Vision, 2000.
[28] Y.-S. Ryu, S.-Y. Oh, Automatic extraction of eye and mouth fields from a face image using eigenfeatures and multilayer perceptrons, Pattern Recogn. 34 (12) (2001) 2459–2466.
[29] M. Balasubramanian, S. Palanivel, V. Ramalingam, Real time face and mouth recognition using radial basis function neural networks, Expert Syst. Appl. 36 (3) (2009) 6879–6888.
[30] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2001, pp. 511–518.
[31] I. Matthews, G. Potamianos, C. Neti, and J. Luettin, A comparison of model and transform-based visual features for audio-visual LVCSR, in: Proc. International Conference on Multimedia and Expo, 2001, no. 2, pp. 2–5.
[32] G. Pomianos, C. Neti, G. Gravier, A. Garg, A.W. Senior, Recent advances in the automatic recognition of audiovisual speech, Proc. IEEE 91 (9) (2003) 1306–1326.
[33] M.T. Chan, HMM-based audio-visual speech recognition integrating geometric- and appearance-based visual features, in: IEEE Fourth Workshop on Multimedia Signal Processing, 2001, pp. 9–14.
[34] G. Potamianos, A. Verma, C. Neti, G. Iyengar, S. Basu, A cascade image transform for speaker independent automatic speechreading, in: Proc International Conference on Multimedia and Expo, vol. II, 2000, pp. 1097–1100.
[35] M. Kass, A. Witkin, D. Terzopoulos, Snakes: active contour models, Int. J. Comput. Vision 1 (4) (1988) 321–331.
[36] T.F. Cootes, D. Cooper, C.J. Taylor, J. Graham, Active shape models – their training and application, Comput. Vis. Image Underst. 61 (1) (1995) 38–59.
[37] G. Potamianos, C. Neti, J. Huang, J. H. Connell, S. Chu, V. Libal, E. Marcheret, N. Haas, J. Jiang, Towards practical deployment of audio-visual speech recognition, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 3, 2004, pp. 777–780.
[38] D.G. Stork, M.E. Hennecke, Speechreading by Humans and Machines, Springer, Berlin, Germany, 1996.
[39] E. Petajan, B. Bischoff, D. Bodoff, N.M. Brooke, An improved automatic lipreading system to enhance speech recognition, in: Proceedings of the SIGCHI Conference on Human factors in Computing Systems, 1988, pp. 19–25.
[40] C. Bregler, H. Hild, S. Manke, A. Waibel, Improving connected letter recognition by lipreading, in: Proceedings of the 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1993, pp. 557–560.
[41] I. Matthews, T.F. Cootes, J.A. Bangham, S. Cox, R. Harvey, Extraction of visual features for lipreading, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2) (2002) 198–213.
[42] D.G. Stork, G. Wolff, E. Levine, Neural network lipreading system for improved speech recognition, in: International Joint Conference on Neural Networks, 1992, pp. 289–295.
[43] M. Gordan, C. Kotropoulos, I. Pitas, A support vector machine-based dynamic network for visual speech recognition applications, EURASIP J. Appl. Signal Process. 2002 (1) (2002) 1248–1259.
[44] P. Yin, I. Essa, J.M. Rehg, Asymmetrically boosted hmm for speech reading, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, 2004, pp. 755–761.
[45] J. Sharp, Microsoft Visual C# 2010 Step by Step, Microsoft Press, Redmond, Washington, 2010.
[46] G. Bradski, A. Kaehler, Learning OpenCV: Computer Vision with the OpenCV Library, O'Reilly Media, 2008.
[47] C. Zhang, Z. Zhang, A Survey of Recent Advances in Face Detection, Microsoft Research, 2010.
[48] R. Lienhart, J. Maydt, An extended set of haar-like features for rapid object detection, in: International Conference on Image Processing, 2002, pp. 900–903.
[49] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. 55 (1) (1997) 119–139.
[50] C. Berry, N. Harte, Region of interest extraction using colour based methods on the cuave database, in: IET Irish Signals and Systems Conference, 2009, pp. 1–6.
[51] A. Albiol, L. Torres, E.J. Delp, Optimum color spaces for skin detection, in: International Conference on Image Processing, vol. 1, 2001, pp. 122–124.
[52] V. Vezhnevets, V. Sazonov, A. Andreeva, A survey on pixel-based skin colour detection techniques, in: Proceedings of Graphicon, 2003, pp. 85–92.
[53] P. Kakumanu, S. Makrogiannis, N. Bourbakis, A survey of skin-color modeling and detection methods, Pattern Recogn. 40 (3) (2007) 1106–1122.
[54] M2VTS Database. <http://www.tele.ucl.ac.be/M2VTS/>.
[55] S. Suzuki, K. Be, Topological structural analysis of digitized binary images by border following, Comput. Vis. Graph. Image Process. 30 (1) (1985) 32–46.
[56] M. de Berg, M. Cheong, O. van Kreveld, M. Overmars, Computational Geometry: Algorithms and Applications, Third ed., Springer-Verlag, 2008, p. 386.
[57] H. Sakoe, S. Chiba, Dynamic programming algorithm optimization for spoken word recognition, IEEE Trans. Acoust. Speech Signal Process. 26 (1) (1978) 43–49.
[58] H. Liu, Study on Lipreading Recognition Based on Computer Vision, in: 2nd International Conference on Information Engineering and Computer Science, vol. 2, no. 1, 2010, pp. 1–4.
[59] M. Müller, Information Retrieval for Music and Motion, Springer, 2007, p. 318.
[60] H. Li, M. Greenspan, Model-based segmentation and recognition of dynamic gestures in continuous video streams, Pattern Recogn. 44 (8) (2011) 1614–1628.
[61] M.Z. Ibrahim, D.J. Mulvaney, Geometry based lip reading system using multi dimension dynamic time warping, in: IEEE International Conference on Visual Communications and Image Processing (VCIP), 2012, pp. 1–6.
[62] P. Kuo, P. Hillman, J. Hannah, Improved lip fitting and tracking for model-based multimedia and coding, in: IEE International Conference on Visual Information Engineering, 2005, pp. 251–258.
[63] N. Eveno, A. Caplier, P. Coulon, Accurate and quasi-automatic lip tracking, IEEE Trans. Circ. Syst. 14 (5) (2004) 706–715.
[64] Y. Yokogawa, N. Funabiki, T. Higashino, M. Oda, Y. Mori, A proposal of improved lip contour extraction method using deformable template matching and its application to dental treatment, Syst. Comput. Jpn. 38 (5) (2007) 80–89.
[65] Y. L. Gillick, S.J. Cox, Some statistical issues in the comparison of speech recognition algorithms, in: Proc. Int. Conf. Acoustics, Speech, Signal Processing, Glasgow, U.K., 1989, pp. 532–535.
[66] M. Gurban, J.-P. Thiran, Information theoretic feature extraction for audio-visual speech recognition, IEEE Trans. Signal Process. 57 (12) (2009) 4765–4776.
[67] E. Benhaim, H. Sahbi, G. Vitte, Designing relevant features for visual speech recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 2420–2424.
[68] G. Papandreou, A. Katsamanis, V. Pitsikalis, P. Maragos, Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition, IEEE Trans. Audio Speech Lang. Process. 17 (3) (2009) 423–435.