Research Paper

A Novel Visual Speech Representation and HMM Classification for Visual Speech Recognition

Dahai Yu,[†1,†2] Ovidiu Ghita,[†1] Alistair Sutherland[†2] and Paul F. Whelan[†1]

This paper presents the development of a novel visual speech recognition (VSR) system based on a new representation that extends the standard viseme concept (that is referred in this paper to as Visual Speech Unit (VSU)) and Hidden Markov Models (HMM). Visemes have been regarded as the smallest visual speech elements in the visual domain and they have been widely applied to model the visual speech, but it is worth noting that they are problematic when applied to the continuous visual speech recognition. To circumvent the problems associated with standard visemes, we propose a new visual speech representation that includes not only the data associated with the articulation of the visemes but also the transitory information between consecutive visemes. To fully evaluate the appropriateness of the proposed visual speech representation, in this paper an extensive set of experiments have been conducted to analyse the performance of the visual speech units when compared with that offered by the standard MPEG-4 visemes. The experimental results indicate that the developed VSR application achieved up to 90% correct recognition when the system has been applied to the identification of 60 classes of VSUs, while the recognition rate for the standard set of MPEG-4 visemes was only in the range 62–72%.

1. Introduction

Automatic Visual Speech Recognition (VSR) plays an important role in the development of many multimedia systems such as audio-visual speech recognition (AVSR)[1], mobile phone applications, human-computer interaction and sign language recognition[2]. Visual speech recognition involves the process of interpreting the visual information contained in a visual speech sequence in order to extract the information necessary to establish communication at perceptual level between humans and computers. The availability of a system that is able to interpret the visual speech is opportune since it can improve the overall accuracy of audio or hand recognition systems when they are used in noisy environments.

The task of solving visual speech recognition using computers proved to be more complex than initially envisioned. Since the first automatic visual speech recognition system introduced by Petajan[7] in 1984, abundant VSR approaches have been reported in the computer vision literature over the last two decades. While the systems reported in the literature have been in general concerned with advancing theoretical solutions to various subtasks associated with the development of VSR systems, this makes their categorization difficult. However, the major trends in the development of VSR can be divided into three distinct categories: feature extraction, visual speech representation and classification. In this regard, the feature extraction techniques that have been applied in the development of VSR systems can be divided into two main categories, shape-based and intensity based. In general, the shape-based feature extraction techniques attempt to identify the lips in the image based either on geometrical templates that encode a standard set of mouth shapes[17] or on the application of active contours[3]. Since these approaches require extensive training to sample the spectrum of mouth shapes, the feature extraction has recently been carried out in the intensity domain. Using this approach, the lips are extracted in each frame based on the colour information and the identified image sub-domain detailing the lips is compressed to obtain a low-dimensional representation.

A detailed review on the research on VSR indicates that numerous methods have been proposed to address the problems of feature extraction and visual speech classification, but limited research has been devoted to the identification of the most discriminative visual speech elements that are able to model the speech process in the continuous visual domain. In this regard, most of the research in the field of visual speech representation was concerned with the identification of the optimal strategy that is able to map the elementary linguistic elements such as phonemes in the visual domain. To this end, the most basic visual speech unit that is able to describe the speech process is the viseme which can be thought of as the counterpart of the phoneme in the visual domain. The viseme representation has attracted substantial research interest and this is mainly motivated by

†1 Vision Systems Group, School of Electronic Engineering, Dublin City University
†2 School of Computing, Dublin City University

the fact that this visual speech strategy requires only a small number of viseme categories to describe the continuous visual speech. While the application of the viseme representation to model the visual speech is appealing due to its simplicity and flexibility, the selection of the most representative viseme categories still is an open research issue. Thus, many investigations in the area of VSR have been conducted using viseme-based speech representations where the number of viseme categories has been varied from 6 to 50 [13]. Although the identification of the optimal viseme representation still is an ongoing research topic, the MPEG-4 viseme set has been the most used representation in the development of VSR systems. This is partially motivated by the fact that the MPEG-4 viseme set has been adopted to facilitate the analysis of the facial dynamics that are encountered during the continuous speech, a process that is closely related to VSR. While the viseme-based representation is intuitive as it can be applied in the generation of phonetic sequences in the visual domain, recent studies have indicated that this visual speech representation has associated several disadvantages when included in the development of practical VSR systems. The most important issues are related to their poor discriminative power, as the visemes have limited visual support and more importantly they are highly sensitive to the lexical context. In other words, the viseme visual context is highly influenced by coarticulation rules that are enforced during the continuous speech process. To address the vulnerability of the standard viseme representation to the lexical context, several works on VSR proposed the introduction of compound speech elements such as bivisemes, trivisemes [18)–23] or visyllables [24]. These compound viseme-based representations shown increased stability with respect to coarticulation rules (lexical context), but it is important to note that the number of categories associated with the biviseme and triviseme representations is substantially larger than the categories associated with the MPEG-4 standard (14 viseme classes). This is one of the main reasons that restricted the application of complex visual speech representation such as visyllables in the implementation of VSR systems and as a result the vast majority of approaches on VSR have attempted to model the visual speech using either biviseme or triviseme representations.

The main goal of this paper is to propose a new visual speech modelling strategy based on Visual Speech Units (VSU) that is able to incorporate the inter-visual context between consecutive visemes. In this regard the VSUs extend the standard viseme representation by including additional lexical context that is represented by the information that sample the transition between consecutive visemes. As opposed to other related implementation based on biviseme or triviseme representations that were evaluated in the context of audiovisual speech recognition [19),21] and text-to-audiovisual speech synthesis [18], where the main problems are located in the process of synchronizing the audio and visual information or generating photo-realistic face animations, the main contribution of this work resides in the development of composite viseme descriptors (VSUs) that are specifically designed to solve video-only speech recognition tasks. It is important to note that the absence of audio information generates a more difficult scenario to perform the localization and recognition of the visual speech descriptors, as these tasks have to be carried out only in the visual domain. Thus, the major objective of this paper is to advance algorithmic solutions for robust feature extraction and to propose a novel strategy that is able to perform the alignment between the trained VSU models and the image data using non-rigid registration techniques. To support the validity of our approach, a large number of experiments are conducted to demonstrate that the inclusion of this new visual speech representation in the development of VSR leads to improved performance when compared with the performance offered by the standard set of MPEG-4 visemes.

This paper is organised as follows. Section 2 details the construction of the Expectation-Maximization Principal Component Analysis (EM-PCA) manifolds and the problems related to their application to VSR tasks. Section 3 provides a detailed analysis of the viseme representation where the main emphasis of the discussion is placed on the analysis of the main limitations associated with this visual speech modelling strategy. Section 4 introduces the speech modelling strategy based on Visual Speech Units (VSUs) and a substantial discussion is provided to illustrate the advantages associated with these new visual speech elements when compared to the standard MPEG-4 viseme categories. In this section additional details related to the registration of the VSU training models and the adopted classification scheme are also included. Section 5 presents the experimental results, while Section 6 concludes this paper with a summary of the findings resulting from our study.

2. Lip Segmentation and EM-PCA Manifold Representation

2.1 Lip Segmentation

To enhance the presence of the skin in the image, the pseudo-hue[5] component is calculated from the RGB representation for each frame in the video sequence. Using this approach the image data can be approximated with a bi-modal distribution and the skin areas defined by pixels characterized by strong red and green components are extracted by applying a thresholding operation, where the threshold value is automatically selected as the local minima between the first and the second peak of the pseudo-hue histogram. The region defining the lips is identified by employing a simple validation procedure that attempts to rank all regions resulting from the application of the thresholding procedure with respect to their position in the image. The images resulting from the lip segmentation procedure are as shown in **Fig. 1**, where the image depicted in Fig. 1 (f) is used as input data to generate the manifold representation. This will be discussed in the next section.

2.2 EM-PCA Manifold Generation

In order to reduce the dimensionality of the data resulting from the lip segmentation process, data compression techniques are applied to extract the lip-features from each frame in the video sequence. To achieve this goal, an Expectation-Maximization Principal Component-Analysis (EM-PCA) scheme is applied to obtain a compact representation for all images resulting from the lip segmentation procedure[6]. The Expectation-Maximization (EM) is a probabilistic framework that is applied to learn the principal components of a dataset using a space partitioning approach. Its main advantage resides in the fact that it does not require to compute the sample covariance matrix as the standard PCA technique and has a complexity limited to *O(knp)* where k is the number of leading eigenvectors to be learned, n is the dimension of the unprocessed data and p defines the number of vectors required for training.

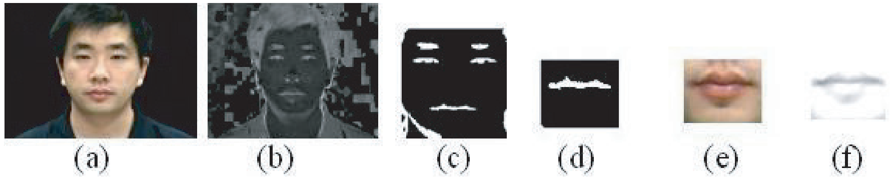

Fig. 1 Lip segmentation process. (a) Original RGB image. (b) Pseudo-hue component calculated from the RGB image shown in (a). (c) Image resulting after thresholding. (d) Image describing the mouth region. (e) ROI extracted from the original image (f) Gray-scale normalized ROI.

EM-PCA is an extension of the standard PCA technique by incorporating the advantages of the EM algorithm in terms of estimating the maximum likelihood values for missing information. This technique has been originally developed by Roweis[6] and its main advantage over the standard PCA is the fact that it is more appropriate to handle large high dimensional datasets especially when dealing with missing data and sparse training sets. The EM-PCA procedure has two distinct stages, the E-step and M-step:

$$E - step : W = (V^T V)^{-1} V^{-1} A \tag{1}$$

$$M - step : V_{new} = A W^T (W W^T)^{-1} \tag{2}$$

where 'W' is the matrix of unknown states, 'V' is the test data vector, 'A' is the observation data and$T$ is the transpose operator. The columns of 'V' span the space of the first k principal components.

As explained in the previous section, the lips regions are segmented in each frame and the appearance of the lips is encoded as a point in a feature space that is obtained by projecting the input data onto the low dimensional space generated by the EM-PCA procedure. The feature points obtained after data projection on the low-dimensional EM-PCA space are joined by a poly-line by ordering the frames in ascending order with respect to time (see **Fig. 2**) to generate the manifold representation. In the implementation detailed in this paper we used only the first three EM-PCA components ($k = 3$) to construct the low-dimensional manifolds since they are able to sample more than 90% of the statistical variance of the images contained in the training stage. In our experiments we have used a training database that consists of 40,000 images.

2.3 Manifold Interpolation

Since the manifolds encode the appearance of the lips in consecutive frames

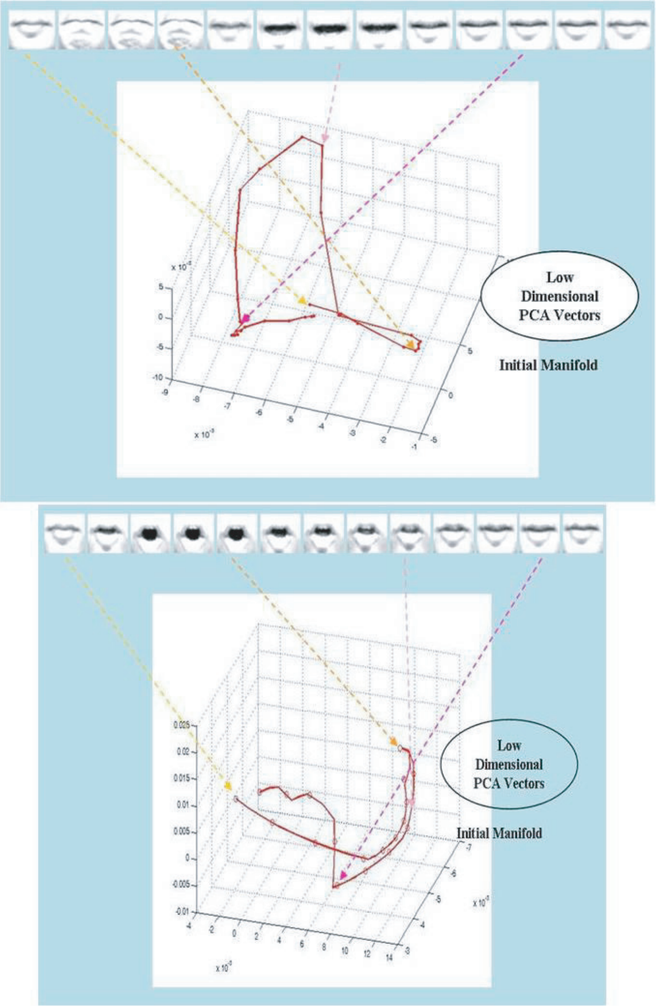

Fig. 2 EM-PCA manifold representation of the word 'Bart' (top) and 'Hot' (bottom). Each feature point of the manifold is obtained by projecting the image data onto the low-dimensional EM-PCA space.

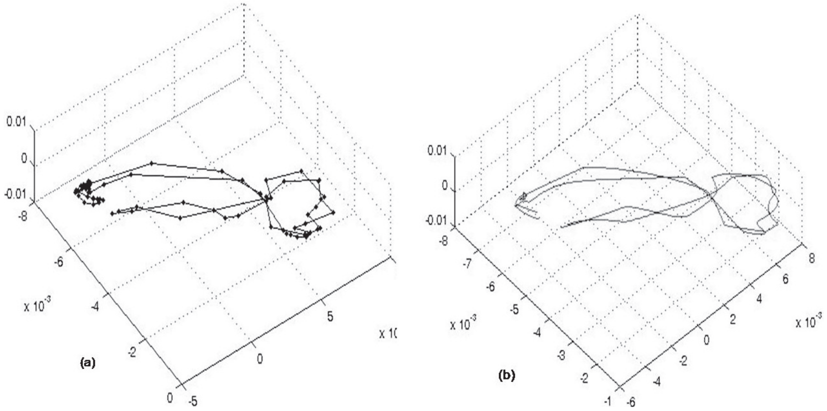

Fig. 3 Manifold representation and interpolation (a) Manifold generated from two image sequences of the word "hook". (b) Manifold interpolation results.

through image compression, the shape of the manifold is strongly related to the words spoken by the speaker and recorded in the input video sequence. **Figure 3** (a) illustrates the manifolds calculated for two independent image sequences describing the same word. Although the video sequences have been generated by two speakers, it can be observed that the shapes of the manifolds are very similar.

While the manifold determined as illustrated in Fig. 3 (a) is defined by a discrete number of points that is given by the number of frames in the video data, this manifold representation is not convenient to be used for classification/recognition purposes since the spoken words may be sampled into a different number of frames that may vary when the video data is generated by different speakers. To address this issue, the feature points that define the manifold are interpolated using a cubic-spline to obtain a continuous representation of the manifold [8]. The manifolds resulting from the interpolation procedure are depicted in Fig. 3 (b). The main issue related to the identification of the speech elements that define the word manifolds is associated with the generation of a visual representation that performs an appropriate phoneme mapping in the visual domain. This problem will be addressed in detail in the next section of this paper.

3. Viseme Representation

3.1 Viseme Background

The basic unit that describes how speech conveys linguistic information is the phoneme. In visual speech, the smallest distinguishable unit in the image domain is called viseme[4),14)]. A viseme can be regarded as a cluster of phonemes and a model for English phoneme-to-viseme mapping has been proposed by Pandzic and Forchheimer[9)].

In 1999, Visser, et al.[10)] developed one of the first viseme-based classification systems where a time-delayed neural network was applied to classify 14 classes of visemes. This work has been further advanced by Foo, et al.[4),16)], where adaptive boosting and HMM classifiers were applied to recognize visual speech visemes. Yau, et al.[11)] followed a different approach when they initially examined the recognition of 3 classes of viseme using motion history image (MHI) segmentation and later they increased the number of visemes up to 9 classes. To describe the lip movements in the temporal domain, 2D spatio-temporal templates (STT) were augmented with features calculated using the discrete wavelet transform and Zernike moments. In their approach HMM classifiers were employed to discriminate between different classes of visemes.

Although there is a reasonably strong consensus about the set of English phonemes, there is less unanimity in regard to the selection of the most representative visemes. Since phonemes and visemes cannot be mapped directly, the total number of visemes is much lower than the number of standard phonemes. In practice, various viseme sets have been proposed with their sizes ranging from 6[12)] to 50 visemes[13)]. Actually this number is by no means the only parameter in assessing the level of sophistication of different schemes applied for viseme categorisation. For example, some approaches propose small viseme sets based on English consonants, while others propose the use of 6 visemes that are obtained by evaluating the discrimination between various mouth shapes (closed, semi-opened and opened mouth shapes). This paper adopts the viseme model established for facial animation by an international object-based video representation standard known as MPEG-4[9)].

From this short literature review, it can be concluded that a viseme is defined as the smallest unit that can be identified using the visual information from the input video data. Using this concept, the word recognition can be approached as a simple time-ordered combination of standard visemes. Although words can be theoretically formed by a combination of standard visemes, in practice viseme identification within words is problematic since different visemes may overlap in the feature space or they may be distorted by the preceding visemes during the continuous speech process.

3.2 Viseme Representation in the EM-PCA Space

In order to evaluate the feasibility of the viseme representation when applied to continuous VSR, a set of MPEG-4 visemes is extracted from input video sequences associated with different words in our database. For instance, frames describing the viseme *[b]* are extracted from words such as 'boat', 'boot', 'batch' etc., while frames describing the viseme *[ch]* are extracted from words such as 'chard', 'choose', 'chocolate' etc.

The feature points that generate the EM-PCA manifold surface describe particular mouth shapes or lip movements and they are manually selected to represent visemes from spoken words. **Figure 4** shows the correspondence between feature points that form the visemes manifolds and the corresponding images that define visemes in the image domain. From this diagram, it can be observed that frames describing standard visemes include three independent states. The first state is the initial state of the viseme; the second state describes the articulation process and the last state models the mouth actions associated with the relaxed state. These frames are projected onto the EM-PCA space and the resulting manifolds are subjected to spline interpolation, as illustrated in **Fig. 5** (a). The feature points for visemes *[b], [u:]* and *[t]* are constructed from video sequences describing the word '*boot' [bu:t]*. By analyzing different instances of the same word '*boot*', a group of features points for visemes *[b], [u:]* and *[t]* is constructed to define each viseme in the manifold representation. These feature points are marked with ellipsoids in the EM-PCA space to indicate the space covered by particular visemes, see Fig. 5 (b). Based on these examples, we can observe that visemes are too small entities to fully characterize the entire word information in the visual domain since the transitions between visemes are not used in the standard viseme-based speech representation. The absence of the inter-viseme

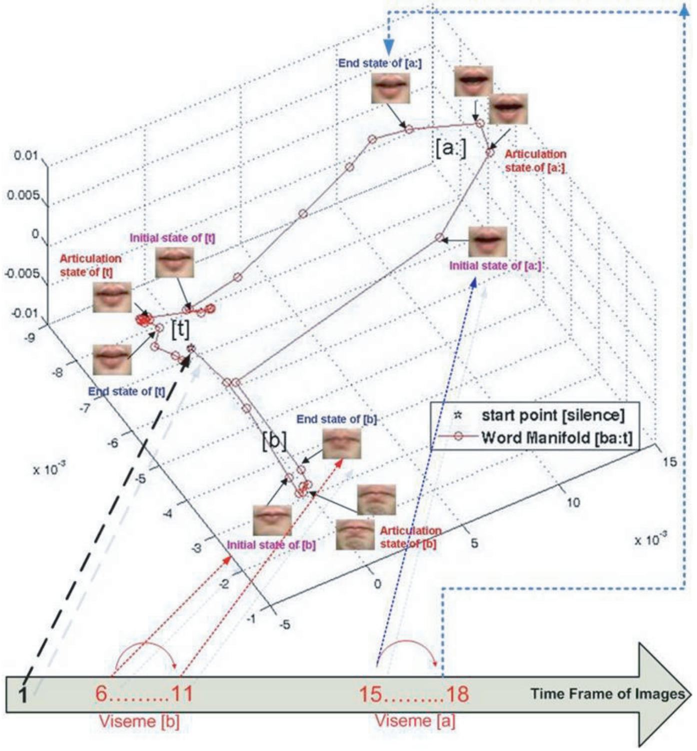

Fig. 4 EM-PCA points generated by images sequence describing the word [ba:t].

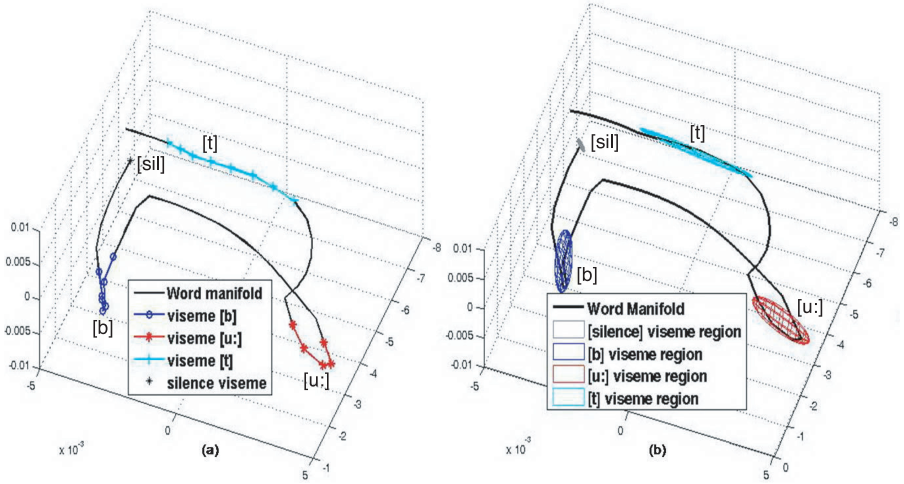

Fig. 5 Viseme representation. (a) EM-PCA feature points associated with visemes [b], [u:] and [t]. (b) The regions in the feature space for visemes [b], [u:] and [t].

transient information is clearly inadequate, as the mouth shapes that describe the transitions between consecutive visemes implement the coarticulation rules that are enforced during the continuous speech process. However, this is not the only deficiency associated with this visual speech representation and a detailed discussion focused on the practical problems encountered when visemes are used to model the visual speech will be provided in the next section.

3.3 Viseme Limitations

As indicated in the previous section, the main shortcoming associated with the viseme representation is given by the fact that large parts of the word manifold (i.e. transitions between visemes) are not used in the recognition process. This approach is inadequate since the inclusion of more instances of the same viseme extracted from different words would necessitate larger regions to describe each viseme in the EM-PCA feature space and this will lead to significant overlaps in the feature space describing different visemes. This problem can be clearly observed in **Fig. 6** where the process of constructing the viseme spaces for two different words ('Bart' and 'chard') is illustrated. As illustrated in Fig. 6, a large region is required to describe the viseme *[a:]* in the feature space of the two different words. Viseme *[d]* (green) in word *[cha:d]* and viseme *[t]* (dark green) in word *[ba:t]* are in the same category of visemes and they also require a large region in the feature space.

Another limitation of the viseme-based representation resides in the fact that due to coarticulation some visemes may be severely distorted and even may disappear in video sequences that describe visually the spoken words. For instance, in the manifolds generated for words 'heart', 'hat', and 'hot' the viseme *[h]* is not visually apparent as no mouth shapes are associated with this viseme when the words are enunciated. This is motivated by the fact that the voicing of the sound 'h' occurs only at glottis and as a result the transitory information that occurs between the viseme *[silence]* and the next viseme *[a:]* for words 'hat' and 'heart' or *[o]* for the word 'hot' may provide a richer source of information when the EM-PCA manifolds are evaluated in the context of VSR.

These simple examples indicate that visemes do not accurately map phonemes in the visual domain and in addition they are subjected to a large degree of distortion when evaluated in continuous speech sequences. We can safely conclude that visemes are too limited to accommodate context-dependent phonetic sequences that are often encountered during the continuous speech process. The limitations associated with the viseme speech representation prompted us to investigate the development of more complex elementary speech units that are able to sample in an elaborate manner the inter-visual context between consecutive visemes.

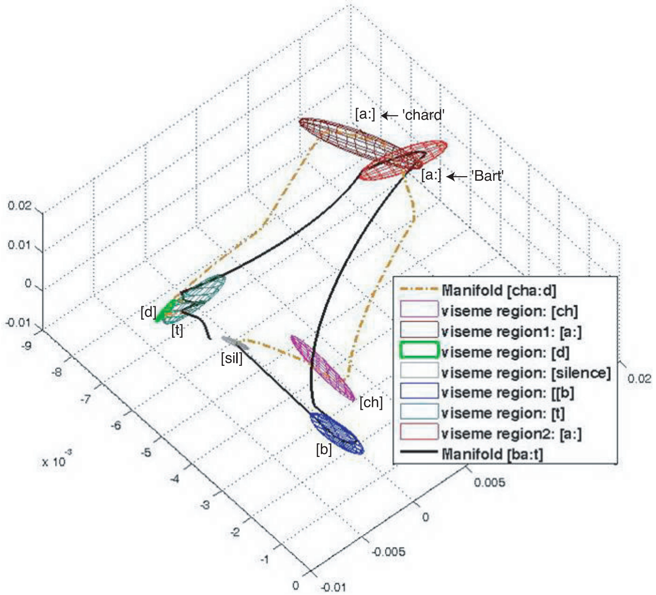

Fig. 6 Viseme feature space constructed for two different words. Word "Bart"-viseme [b], [a:] and [t]. Word "chard" – visemes [ch], [a:] and [d].

4. Visual Speech Units

4.1 Visual Speech Units Modelling

The main aim of this paper is to introduce a new representation called Visual Speech Unit (VSU) that includes not only the data associated with the articulation of standard visemes, but also the transitory information between consecutive visemes. Each VSU is manually constructed from the word manifolds and it has three distinct states: (a) articulation of the first viseme, (b) transition to the next viseme, (c) articulation of the next viseme. The principle behind this new visual speech representation can be observed in **Fig. 7** where a number of prototype examples of VSUs are shown.

4.2 Visual Speech Units Training

As mentioned before, the construction of VSUs is based on adjacent "visible" visemes that can be identified in the word manifolds (the visible visemes describe the articulation process of lip movements that can be mapped in the visual domain). In the manifold representation, the visible visemes are represented as a unique region in the EM-PCA feature space. Using this approach, the VSUs associated with the word 'boot' are: *[silence-b]*, *[b-u:]* and *[u:-t]*, they are displayed in **Fig. 8** (a).

To apply the VSU representation to visual speech recognition it is necessary to construct a mean model for each class of VSU. To facilitate this process, the interpolated word manifolds are re-sampled uniformly into a fixed number of feature-points. In order to generate standard VSUs manifolds for training and recognition tasks, the re-sampling procedure will generate a pre-defined number of keypoints that are equally distanced on the interpolated manifold surface. We

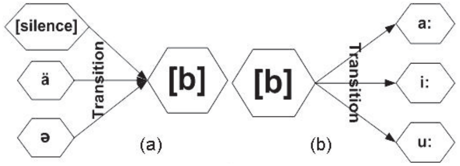

Fig. 7 Visual Speech Unit examples. (a) VSU prototypes: [silence-b], [ä-b] and [æ-b]. (b) VSU prototypes: [b-a:], [b-i] and [b-u:].

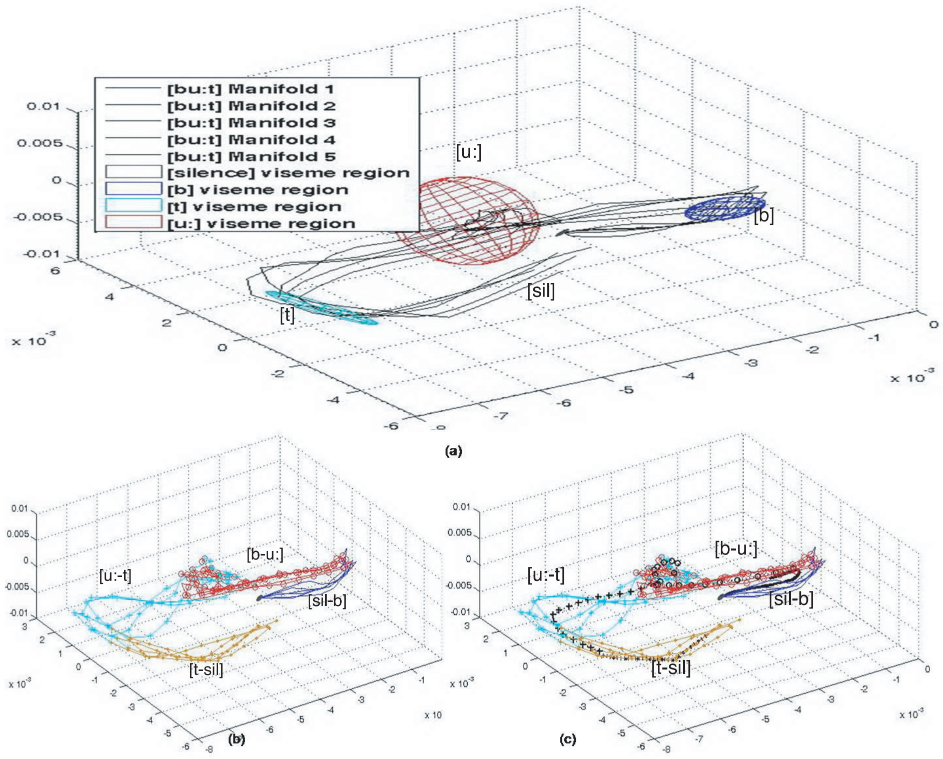

Fig. 8 The calculation of the VSU mean models. (a) Five manifolds of the word [bu:t] (black line), four visible visemes :*[silence]* (gray), *[b]* (blue), *[u:]* (red) and *[t]* (cyan). (b) The VSU manifolds extracted and re-sampled: *[silence - b]* (blue), *[b-u:]* (red), *[u:-t]* (cyan) and *[t-silence]* (yellow). (c) The mean model for all VSUs are marked in black in the diagram *[silence-b]*(black line), *[b-u:]* (black circles), *[u:-t]* (black cross) and *[t-silence]* (black dot).

have determined through experimentation that optimal results are achieved when the interpolated manifolds are uniformly resampled using 50 keypoints. This re-sampling procedure ensures the identification of a standard set of features (keypoints) as illustrated in Fig. 8 (b).

The keypoints for each VSU are manually extracted from different instances of the same word and they are used to train the HMM classifiers and to calculate the mean model. The procedure applied to determine the mean models for all VSUs contained in the word 'boot' is illustrated in Fig. 8 (c). In the implementation presented in this paper, to minimize the class overlap it has been trained one HMM classifier for each VSU class. The procedure applied to train the HMM classifiers is depicted in **Fig. 9**.

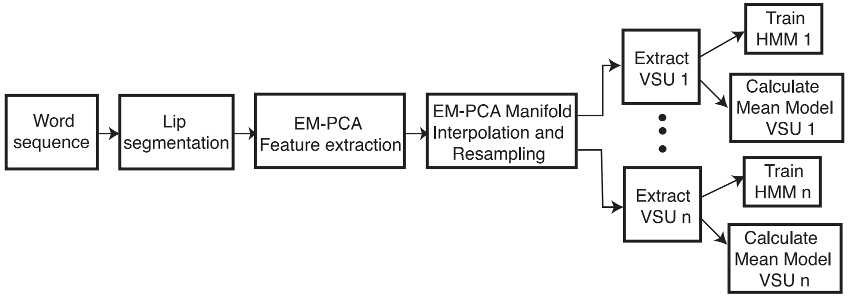
Word sequence
Lip segmentation
EM-PCA Feature extraction
EM-PCA Manifold Interpolation and Resampling
Extract VSU 1
Extract VSU n
Train HMM 1
Calculate Mean Model VSU 1
Train HMM n
Calculate Mean Model VSU n

Fig. 9 The VSU training process.

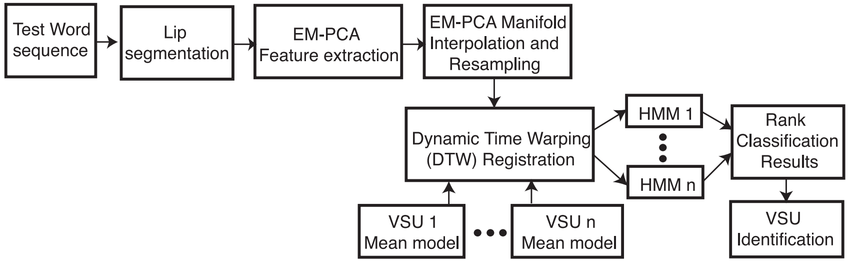
Test Word sequence
Lip segmentation
EM-PCA Feature extraction
EM-PCA Manifold Interpolation and Resampling
Dynamic Time Warping (DTW) Registration
VSU 1 Mean model
VSU n Mean model
HMM 1
HMM n
Rank Classification Results
VSU Identification

Fig. 10 The VSU classification process.

4.3 Registration between VSU and Word Manifolds

The VSU registration and classification process is outlined in **Fig. 10**. The VSU recognition is carried out at word level and it can be viewed as a competitive process where all VSU mean models are registered to the interpolated manifold that is calculated from the input video sequence. In this fashion, we attempt to divide the word manifold into a number of consecutive sections, where each section is compared against the mean models of all VSUs stored in the database. To achieve this, we need to register the VSU mean models with the surface of the word manifold. In this work the registration between VSU mean models and the surface of the word manifolds is carried out using the Dynamic Time Warping (DTW) algorithm. DTW is a simple technique that has been commonly used in the development of VSR systems to determine the similarity between time series and to find corresponding regions between two time series of different lengths [15]. Let X and Y be two time series, of lengths |X| and |Y|, where $W = w_1, w_2, \ldots, w_K$ is the warp path $(\max(|X|, |Y|) \leq K < |X| + |Y|)$, K is the length of the warp path, $w_k = (i, j)$ is the kth element of the path, i is the index for time series X and j is an index for time series Y. The optimal warp path is calculated by minimizing the fitting cost between the two time series as follows:

$$Dist(W) = \sum_{k=1}^{K} Dist(w_{ki}, w_{kj}) \tag{3}$$

where $Dist(W)$ is the distance (typically the Euclidean distance) associated with the warp path W, and $Dist(w_{ki}, w_{kj})$ is the distance between two data points with indexes i and j. The warp path must start at the beginning of each time series and finish at the end of both time series. This ensures that every element of each time series is used in the calculation of the warp path.

The VSU recognition process is implemented as a two-step approach. In the first step we need to register the VSU mean models to the word manifold using DTW, while in the second step we measure the matching cost between the VSU mean models and the registered section of the manifold using HMM classification. We have adopted this strategy to avoid the identification of the maximum likelihood estimate of the parameters of the HMM by performing an exhaustive search over the entire manifold data. This approach is also motivated by the intuitive observation that the recognition of the elementary visual speech units (VSUs) is based on a sequential process. The proposed two-step recognition procedure is applied for all VSUs contained in the database and the complete registration process of the word 'chard' is illustrated in **Fig. 11**.

4.4 HMM Classification

The lips motions associated with VSUs can be partitioned into three HMM states using one Gaussian mixture per state and a diagonal covariance matrix. The first state describes the articulation of the first viseme of the VSU. The second state is defined by the transition to the next viseme, while the third state is the articulation of the second viseme. **Figure 12** (a) illustrates the partition of the VSU into a sequence of three hidden states.

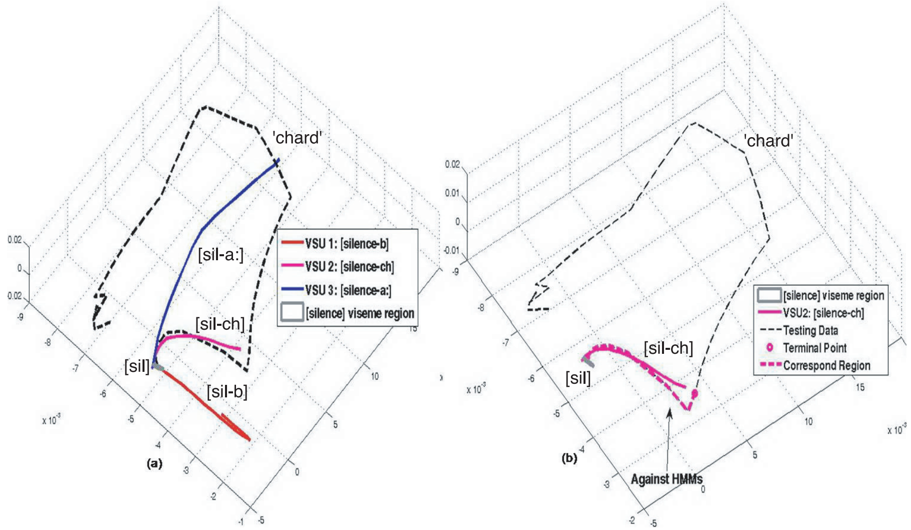

Fig. 11 VSU registration and classification. (a) The registration of three classes of the VSU Class 1: [silence-b] (red line); Class 2: [silence-ch] (purple line); Class 3: [silence-a:] (blue line) to the word manifold (black dotted line). (b) Registration between the [silence-ch] VSU mean model and the word manifold. The [silence-ch] VSU mean model achieved the best matching cost (evaluated using a three-state HMM classification).

In the implementation detailed in this paper, the unknown HMM parameters are iteratively estimated based on the training samples using a Baum-Welch algorithm. We have constructed one HMM classifier for each class of VSU and one HMM classifier for each viseme as well. Each trained HMM estimates the likelihood between the registered section of the word manifold and the VSU mean models stored in the database. The HMM classifier that returns the highest likelihood will map the input visual speech to a particular class in the database. Figure 12 (b) shows one example where the mouth shapes associated with the [b-a:] VSU are modelled using three-state HMMs.

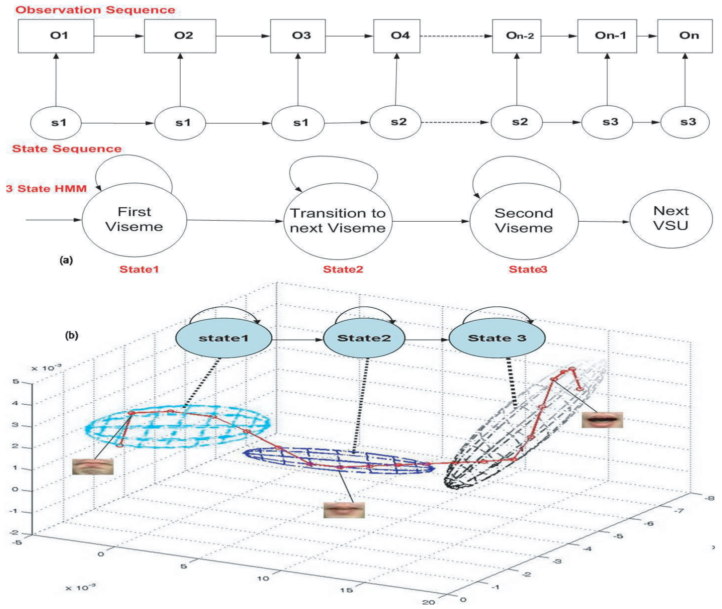

Fig. 12 (a) HMM topology for VSU. (b) The partition of the manifold of the VSU - [b-a:] using HMMs.

5. Experimental Results

For evaluation purposes it has been created a database that is generated by two speakers. This database consists of 50 words (**Table 1**) where each word is spoken 10 times by speaker one and 20 words where each word is spoken 6 times by speaker two. The words were selected to generate a complex testing scenario that allows the evaluation of the proposed visual speech modelling strategy based on VSUs when included in the development of a VSR system. In this regard, in our database we have included simple words such as 'boat', 'heart', 'check', etc. and more complex words such as 'Barbie', 'Hoover', 'bookman', 'chocolate', etc. The words contained in the database facilitate the evaluation of the viseme and VSU speech representations in the presence of contextual viseme distortions, which is one of the main goals of this paper. In our study we have conducted the experiments to evaluate the recognition rate when 12 classes of visemes [9] and 60 classes of VSUs (**Table 2** and **Table 3**) are used as speech elements. As indicated in Table 3, the VSU classes are divided into three distinct groups: Group 1 (First VSUs) contains the VSUs that start with the state *[silence]*, Group 2

Table 1 Word database.

Speaker	Words
1	Bart, boat, beat, bet, bird, boot, barbie, book, beef, barge, birch, bookman, batch bobby, beefalo beautiful, before, heart, hot, heat, hat, hook, harp, hobby, hoover, half, home, chard, choose, cheat, check, charge, cheap, channel, charming, chocolate, chief, wart, zart, fast, banana, January, truth, part, put, mart, mood, I, bar, card.
2	Bart, boat, beat, boot, heart, hot, heat, hook, charge, choose, cheat, check, wart, zart, fat, bar, art, ill, oat, fool.

Table 2 The set of MPEG-4 visemes.

Viseme Number	Phonemes	Example Words	No. of samples
1	[b], [p], [m]	boot, part, mart	300
2	[s], [z]	zart, fast	30
3	[ch], [dZ]	chard, charge	150
4	[f], [v]	fast, hoover	80
5	[I]	beat, heat	130
6	[A:]	Bart, chard,	250
7	[e]	hat, bet	130
8	[O]	boat, hot	100
9	[U]	hook, choose	80
10	[t, d]	boot, bird,	190
11	[h, k, g]	card, hook,	130
12	[n,l]	banana	20

Table 3 60 classes of visual speech units.

VSU Groups	Numbers	VSUs
Group 1:	9	[silence-b], [silence-ch], [silence-z], [silence-f], [silence-a:], [silence-o], [silence-i:], [silence-e], [silence-u:]
Group 2:	16	[a:-silence], [o-silence], [æu:-silence], [u-silence], [k-silence], [i:-silence], [ch-silence], [f-silence], [m-silence], [ing-silence], [æ-silence], [p-silence], [et-silence], [g-silence], [s-silence], [ë-silence]
Group 3:	35	[b-a:], [b-o:], [b-i:], [b-u:], [b- ë], [b-æ], [a:-t], [a:-b], [a:-f], [a:-g], [a:-ch], [o-b], [o-t], [o-k], [i:-f], [i:-p], [i:-t], [u:-t], [u:-k], [u:-f], [æ-t], [f-ë:],[f-o], [k-m], [f-a:], [w-a:], [z-a:], [a:-t], [ë:-n], [ë:-ch], [n-a:], [a:-n], [ch-a:], [ch-u:], [ch-i:]

(Last VSUs) contains the VSUs that end with *[silence]*, while Group 3 contains the 'middle' VSUs (middle VSUs are the speech elements that are not defined by transitions *[silence - viseme]* and *[viseme - silence]*). The main motivation behind the decision to segregate the VSUs classes into three groups is twofold. First, this approach allows us to evaluate the drop in classification accuracy associated with the middle VSUs when compared with accuracies achieved by the VSUs contained in Groups 1 and 2. The drop in classification accuracy for middle VSUs is expected since the coarticulation rules distort more the visual context associated with this category of VSUs than the visual context associated with the VSUs contained in Groups 1 and 2. Second, this approach opens the possibility to assess in greater details the performance of the proposed VSR system when evaluated on data generated by multiple users.

Although 60 classes of VSUs do not cover the entire spectrum of possible viseme combinations, they allow us to conduct detailed experiments to assess the performance of the proposed visual speech representation when compared to that achieved by the standard MPEG-4 visemes. To avoid the bias in the training stage, as the number of words generated by the two speakers is different, we elected to independently train the HMM classifiers for each speaker. The segregation of the training stage was also motivated by the substantial differences in appearance between similar categories of mouth shapes that were encountered in the image data generated by the two speakers (one male and one female). It is useful to note that the database generated by the second speaker is defined by a smaller number of words when compared to that generated by the first speaker. As a result not all of the 60 classes of VSUs listed in Table 3 could be generated and tested using the database produced by the second speaker. Thus, when the proposed system has been evaluated on data generated by the second speaker only a subset of visemes (10 classes) and VSUs (30 classes) could be used in the experimental activity.

The experimental tests were divided into two sets. The first tests were conducted to evaluate the classification accuracy when standard MPEG-4 visemes and VSUs are employed as speech elements and the number of words in the database is incrementally increased. The classification results for speaker one are depicted in **Fig. 13** (a) and for speaker two are depicted in Fig. 13 (b). Based on these experimental results, it is noticed that the correct identification of the visemes in the input video sequence drops significantly with the increase in the number of words in the database. Conversely, the recognition rate for VSUs suffers a minor reduction with the increase in the size of the database.

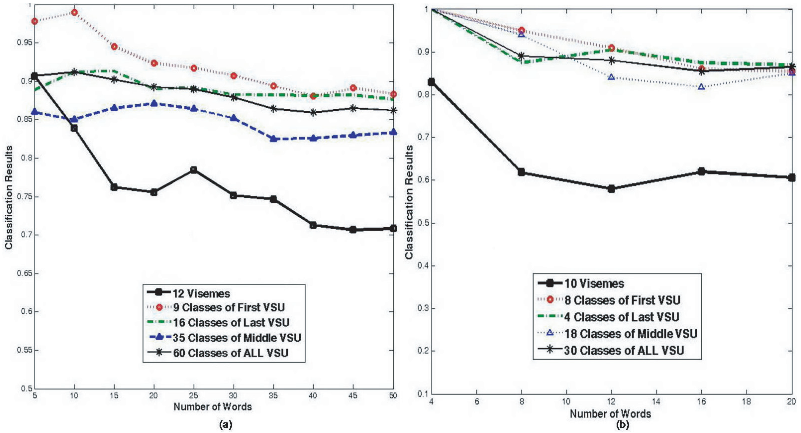

Fig. 13 Viseme vs. VSU classification. (a) Speaker one. (b) Speaker two.

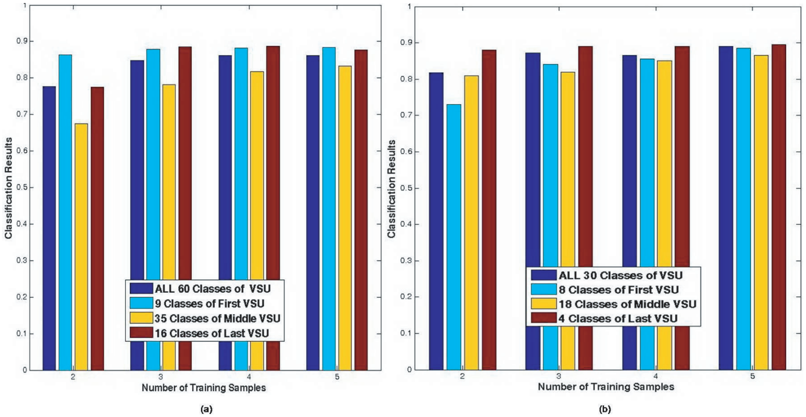

Fig. 14 Visual Speech Unit classification with respect to the number of training examples. (a) Speaker one; (b) Speaker two.

The aim of the second set of experiments is to evaluate the performance of the VSU recognition with respect to the number of samples used to train the HMM classifiers. As expected, the recognition rate is higher when the number of samples used in the training stage is increased (**Fig. 14**). It can be also observed that the recognition rate of Group 3 (middle VSUs) is lower than the recognition rate for Groups 1 and 2. This is explained by the fact that the VSUs contained in Groups 1 and 2 start or end with *[silence]* and this state can be precisely located in the word manifold. Another motivation is given by the increased intra-class variability associated with the middle VSUs, as the number of classes for this category of VSU is higher than the number of VSU classes encompassed by the Groups 1 and 2.

6. Conclusions

In this paper we have described the development of a VSR system where the main emphasis was placed on the evaluation of the discriminative power offered by a new visual speech representation that is referred to as a Visual Speech Unit (VSU). The VSU extends the standard viseme concept by including in this new representation the transition information between consecutive visemes.

To evaluate the classification accuracy obtained for the proposed visual speech representation, we have constructed 60 classes of VSUs that are generated by two speakers and we quantified their performance when compared with that offered by the standard set of MPEG-4 visemes. The experimental results presented in this paper indicated that the recognition rate for VSUs is significantly higher than that obtained for MPEG-4 visemes.

In our future studies, we will extend the number of VSU classes and test the developed VSR system on more complex word databases that are generated by a large number of speakers. Future research will be also concerned with the inclusion of the VSU based visual speech recognition in the implementation of a robust sign language gesture recognition system.

References

1) Potamianos, G., Neti, C., Gravier, G., Garg, A. and Senior, A.W.: Recent advances in the automatic recognition of Audio-Visual Speech, *Proc. IEEE*, Vol.91, No.9, pp.1306–1326 (2003).

2) Shamaie, A. and Sutherland, A.: Accurate Recognition of Large Number of Hand Gestures, *Iranian Conference on Machine Vision and Image Processing*, pp.308–317, ICMVIP Press (2003).

3) Luettin, J., Thacker, N.A. and Beet, S.W.: Active Shape Models for Visual Speech Feature Extraction, Univ. of Sheffield, U.K., Electronic System Group Report (1995).

4) Dong, L., Foo, S.W. and Lian, Y.: A two-channel training algorithm for Hidden Markov Model and its application to lip reading, *EURASIP Journal on Applied Signal Processing*, pp.1382–1399 (2005).

5) Eveno, N., Caplier, A. and Coulon, P.: A new color transformation for lips segmentation, *4th Workshop on Multimedia Signal Processing*, Cannes, France, pp.3–8, IEEE Press (2001).

6) Roweis, S.: EM algorithms for PCA and SPCA, *Advances in Neural Information Processing Systems*, Vol.10, pp.626–632 (1998).

7) Petajan, E.D.: Automatic lipreading to enhance speech recognition, Ph.D. dissertation, Univ. of Illinois, Urbana-Champaign (1984).

8) Yu, D., Ghita, O., Sutherland, A. and Whelan, P.F.: A New Manifold Representation for Visual Speech Recognition, *12th International Conference on Computer Analysis of Images and Patterns*, Vienna, Austria, pp.374–382, LNCS Press (2007).

9) Pandzic, I.S. and Forchheimer, R. (Eds.): *MPEG-4 Facial Animation — The Standard, Implementation and Applications*, John Wiley & Sons Ltd, ISBN 0-470-84465-5 (2002).

10) Visser, M., Poel, M. and Nijholt, A.: Classifying Visemes for Automatic Lipreading, *Proc. 2nd International Workshop on Text, Speech and Dialogue* (*TSD' 99*), *LNAI 1692*, pp.349–352 (1999).

11) Yau, W., Kumar, D.K., Arjunan, S.P. and Kumar, S.: Visual Speech Recognition Using Image Moments and Multi-resolution Wavelet Images, *Computer Graphics, Imaging and Visualisation*, pp.194–199 (2006).

12) Leszczynski, M. and Skarbek, W.: Viseme Recognition — A Comparative Study, *Conference on Advanced Video and Signal Based Surveillance*, pp.287–292, IEEE Press (2005).

13) Scott, K.C., Kagels, D.S., Watson, S.H., Rom, H., Wright, J.R., Lee, M. and Hussey, K.J.: Synthesis of speaker facial movement to match selected speech sequences, *5th Australian Conference on Speech Science and Technology* (1994).

14) Potamianos, G., Neti, C., Huang, J., Connell, J.H., Chu, S., Libal, V., Marcheret, E., Haas, N. and Jiang, J.: Towards Practical Deployment of Audio-Visual Speech Recognition, *International Conference on Acoustics, Speech and Signal Processing*, Vol.3, pp.777–780, IEEE Press (2004).

15) Ratanamahatana, C.A. and Keogh, E.: Everything you know about Dynamic Time Warping is wrong, *3rd SIGKDD Workshop on Mining Temporal and Sequential Data* (2004).

16) Foo, S.W. and Dong, L.: Recognition of visual speech elements using Hidden Markov Models, *Proc. IEEE Pacific Rim Conference on Multimedia*, pp.607–614 (2002).

17) Silveira, L.G., Facon, J. and Borges, D.L.: Visual speech recognition: A solution from feature extraction to words classification, *16th Brazilian Symposium on Computer Graphics and Image Processing*, pp.399–405 (2003).

18) Ezzat, T. and Poggio, T.: Visual speech synthesis by morphing visemes, *International Journal of Computer Vision*, Vol.38, pp.45–57 (2000).

19) Hazen, T.J.: Visual model structures and synchrony constraints for audio-visual speech recognition, *IEEE Transactions on Audio, Speech and Language Processing*, Vol.14, pp.1082–1089 (2006).

20) Lee, S. and Yook, D.: Viseme recognition experiment using context dependent Hidden Markov Models, *Proc. 3rd International Conference on Intelligent Data Engineering and Automated Learning*, Vol.2412, pp.557–561, LCNS Press (2002).

21) Hazen, T.J., Saenko, K., La, C. and Glass, J.R.: A segment-based audio-visual speech recognizer: Data collection, development and initial experiments, *Proc. 6th International Conference on Multimodal Interfaces* (*ICMI'04*), pp.235–242 (2004).

22) Bregler, C., Covell, M. and Slaney, M.: Video rewrite: Driving visual speech with audio, *Proc. 24th Annual Conference on Computer Graphics and Interactive Techniques*, pp.353–360 (1997).

23) Bregler, C., Omohundro, S.M., Covell, M., Slaney, M., Ahmad, S., Forsyth, D.A. and Feldman, J.A.: Probabilistic Models of Verbal and Body Gestures, Computer Vision in Man-Machine Iterfaces, Cipolla, R. and Pentland, A. (Eds.), Cambridge University Press (1998).

24) Kshirsagar, S. and Thalmann, N.M.: Visyllable based speech animation, *Proc. EUROGRAPHICS 2003*, pp.631–640 (2003).

(Received March 23, 2009)
(Accepted December 28, 2009)
(Released March 11, 2010)

(Communicated by *Toshikazu Wada*)

Dahai Yu received his B.Eng. and Ph.D. degrees from Dublin City University. His research interests are in the areas of image processing, machine learning and visual speech recognition. Currently, Dr. Yu is with Ericsson Ltd., Ireland.

Ovidiu Ghita received his B.E. and M.E. degrees in Electrical Engineering from Transilvania University Brasov, Romania and his Ph.D. degree from Dublin City University, Ireland. From 1994 to 1996 he was an Assistant Lecturer in the Department of Electrical Engineering at Transilvania University. Since then he has been a member of the Vision Systems Group at Dublin City University (DCU) and currently he holds a position of DCU-Research Fellow. Dr. Ghita has authored and co-authored over 80 peer-reviewed research papers in areas of instrumentation, range acquisition, machine vision, texture analysis and medical imaging.

Alistair Sutherland is a Lecturer in Computing at DCU. He has worked for 15 years in the area of sign language recognition from real-time video. He is developing techniques based on manifolds and multi-scale analysis. He has previous industrial experience with Hitachi Dublin Laboratory. He has a Ph.D. in image processing from the University of Edinburgh.

Paul F. Whelan received his B.Eng. (Hons) degree from NIHED, M.Eng. from the University of Limerick, and his Ph.D. (Computer Vision) from the University of Wales, Cardiff, UK (Cardiff University). During the period 1985–1990 he was employed by Industrial and Scientific Imaging Ltd and later Westinghouse (WESL), where he was involved in the research and development of high speed computer vision systems. He was appointed to the School of Electronic Engineering, Dublin City University (DCU) in 1990 and is currently Professor of Computer Vision (Personal Chair). Prof. Whelan founded the Vision Systems Group in 1990 and the Centre for Image Processing and Analysis in 2006 and currently serves as its director. As well as publishing over 150 peer reviewed papers, Prof. Whelan has co-authored 2 monographs and co-edited 3 books. His research interests include image segmentation, and its associated quantitative analysis with applications in computer/machine vision and medical imaging. He is a Senior Member of the IEEE, a Chartered Engineer and a member of the IET and IAPR. He served as a member of the governing board (1998–2007) of the International Association for Pattern Recognition (IAPR), a member of the International Federation of Classification Societies (IFCS) council and President (1998–2007) of the Irish Pattern Recognition and Classification Society (IPRCS).