USING A FREE-PARTS REPRESENTATION FOR VISUAL SPEECH RECOGNITION

Patrick Lucey+, Simon Lucey and Sridha Sridharan+*

+Speech, Audio, Image and Video Research Laboratory
Queensland University of Technology
GPO Box 2424, Brisbane 4001, Australia
*Advanced Multimedia Processing Laboratory
Department of Electrical and Computer Engineering
Carnegie Mellon University, Pittsburgh PA 15213, USA

+p.lucey@qut.edu.au, *slucey@ieee.org, +s.sridharan@qut.edu.au

ABSTRACT

Motivated by the success of free-parts based representations in face recognition, we have attempted to address some of the problems associated with applying such a philosophy to the task of speaker-independent visual speech recognition. A major problem with canonical area-based approaches in automatic visual speech recognition is the dependence these approaches have on locating and tracking the speaker's region of interest (ROI) correctly. By employing a free-parts representation, we assume that the position/structure of patches within the mouth image can be relaxed so they can "freely" move to varying extents, hence reducing the influence of the front-end effect. In this paper, we show that by using a free-parts representation we gain some robustness against the problem of ROI localisation and tracking compared to current area-based feature extraction techniques such as the discrete cosine transform (DCT). Also in this paper, we expose the importance of representation for the task of visual speech recognition highlighted by the poor results current representations yield.

1. INTRODUCTION

It is largely agreed upon that the majority of visual speech information stems from a subject's mouth [1]. As a result, a large proportion of the work that has been conducted in automatic visual speech reading has been towards the goal of finding a suitable mouth representation for recognition purposes. The more discriminant and compact the mouth representation, generally the easier the recognition task. From literature [2], visual (mouth) features can be categorized into two types, namely: contour and area based representations. Area based representations are concerned with transforming the whole region of interest (ROI) mouth pixel intensity image into a meaningful feature vector(s). Contour based representations, like those proposed in [3], [4] and [5] are concerned with parametrically atomizing the mouth, based on a priori knowledge of the components of the mouth (i.e. outer and inner labial contour, tongue, teeth, etc.). Although there has been much research in the field of visual feature extraction, it is not clear which approach is best for solving the problem of visual speech reading, especially in visually challenging environments as highlighted in the review paper by Potamianos et al. [6]. However, Potamianos et al. mention that the area-based approaches have the advantages over contour-based as their use is well motivated by human perception of the mouth region as well as being computationally efficient which makes it tractable for real-time implementation. This finding was also backed by their earlier work in [7], in which they found that area-based representations obtained superior performance as well as being more robust to visual noise and compression artifacts. For these reasons, area-based representations were used for this work.

However, a major problem with canonical area-based approaches in visual speech recognition is the dependence these approaches have on locating and tracking the speaker's ROI correctly. In visual speech recognition, if ROI is not tracked accurately, it is intuitive that the performance of the overall system will degrade severely. This is known as *the front-end effect*. This is illustrated in Figure 1, where there are two identical mouth images, the top tracked accurately and the bottom one tracked inaccurately. In this paper, we propose a novel area-based representation of the mouth, which we refer to as *free-parts*, based on recent success this representation has enjoyed in the task of face recognition [8], in an aid to reduce the dependence current area-based approaches have on accurate ROI detection and tracking.

Much improvement has been noted in speech recognition literature [9] by relaxing temporal structure in the speech

Proceedings of the Digital Imaging Computing: Techniques and Applications (DICTA 2005)
0-7695-2467-2/05 $20.00 © 2005 IEEE

Fig. 1. Comparison of ROI tracking variabilities, the top being tracked correctly and the bottom not.

signal (e.g. Gaussian mixture models (GMMs), Hidden Markov Models (HMMs)) when compared to more temporally rigid models (e.g. Dynamic time warping (DTW)). An advantage of dealing with models based on low-dimensional distributions (e.g. GMMs, HMMs) over models based on single points existing in a high-dimensional space (e.g. DTWs) is their inherent ability to generalize. This generalization stems from the increase in the number of training observations and a decrease in the dimensionality of these observations; both being caused from the relaxation of structural constraints in the signal. Extending this idea to the spatial domain, we propose the employment of a free-parts representation that relaxes both spatial and temporal structure in the visual signal.

Free-parts representations assume that the position/structure of patches within the mouth image can be relaxed so they can "freely" move to varying extents. An example of this structural relaxation can be seen in Figure 2. The relaxation of structure in the mouth has the major benefit of obtaining a "distribution" instead of a "point" for each visual speech frame. By utilising a distribution instead of a point structure, we believe that the dependence on the detection and tracking of speaker's mouth or region of interest (ROI) is reduced. In this paper, we investigate whether a area-based free-parts representation can improve robustness to poor ROI detection and tracking or not.

2. APPROACH

For this work, we compared the speaker-independent visual speech recognition performances of a widely used area-based monolithic representations to the area-based free-parts representation on both accurately and poorly tracked ROI's, as shown in Figure 1. Training and evaluation visual speech was taken from the Clemson University, *CUAVE*, audio-visual database [10]. The CUAVE database was selected as it is presently the only common audio-visual database which is available for all universities to use. This is important for benchmarking and comparison purposes. The CUAVE

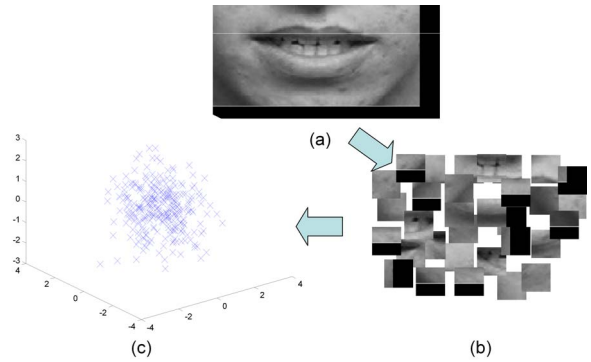

Fig. 2. Depiction of the process of structural collapse of the mouth, from an (a) single monolithic mouth image, through the (b) structural relaxation process (i.e. removing positional information); to finally be left with (c) a cloud of free-parts observations describing the subject's mouth.

database consists of two major sections, one of individual speakers and one of speakers pairs. For this study, only the stationary connected-digit string section of the individual speakers were used. The stationary connected-digit string section of the database consisted of each of the 36 individual speakers uttering the connected digits "zero" to "nine" a total of 5 times each. The 36 individual speakers were divided arbitrarily into a set of 18 training speakers and 18 different test talkers for a completely speaker-independent grouping as per the experiments conducted by Patterson et al. [11].

The task of visual speech recognition is summarised in Figure 3. Each block of the system is explained in the following subsections, detailing how each was implemented for the experiments.

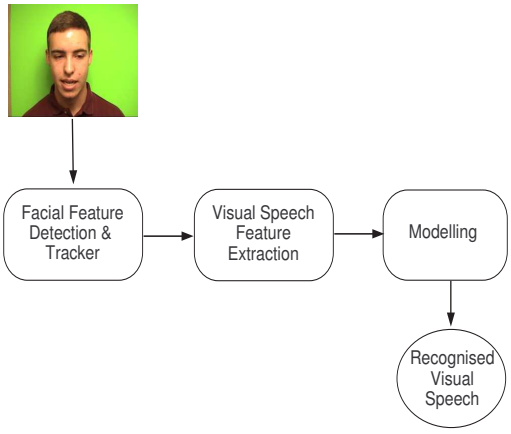

Fig. 3. Block diagram of visual speech recogniser

2.1. Facial Feature Detection and Tracking

Before the visual speech features can be extracted, the ROI has to be detected and tracked. In this study, this consisted of three stages; face location, eye location and lip location. As shown in Figure 4, each stage was used to help form a search region for the next stage.

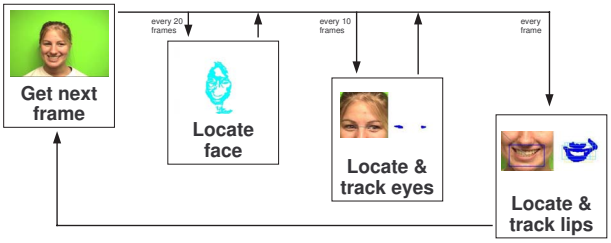

Fig. 4. Overview of lip tracking system.

2.1.1. *Face Location*

Before face location was performed on the videos, 10 manually selected skin points for each speaker are used to form thresholds for the red, green and blue (r, g, b) values in colour-space for skin segmentation. The thresholds for each colour-space were calculated from the skin points as

$$\mu_c - \sigma_c \leq p_c \leq \mu_c + \sigma_c, \tag{1}$$

Where $c \in \{r, g, b\}$, μ_c and σ_c are the mean and standard deviation of the 10 points in colour-space c, and p_c is the value of the pixel being thresholded in colour-space c.

Once the thresholds were calculated, they were used for skin segmentation of the video to generate a bounding box of the face region within the frames every 20 frames, and this face location was remembered in the intermediate frames.

2.1.2. *Eye Location and Tracking*

When transformed into $YCbCr$ space, the eye region of face images exhibit a high concentration of blue-chrominance, and a low concentration of red-chrominance. Therefore eye detection can be done in the $Cr - Cb$ space with reasonable results. However, eyebrows often appear as false positives and can degrade results. To remove the influence of eyebrows the $Cr - Cb$ image can be shifted vertically and subtracted from the original $Cr - Cb$ image. This will cancel the eyebrow minima by subtracting the eye minima, whereas the eye minima will be subtracted by the high values in the skin region and receive a large negative value suitable for thresholding [12].

To locate the eyes from the face region from the previous stage, the top half of the face region was designated as the eye search-area, which was then searched using the shifted $Cr - Cb$ algorithm for the eye locations. The possible eye candidates were searched for two points that were not too large, too close horizontally, and not too distant vertically. Finally the two candidates which had the largest horizontal distance were chosen to be the eye locations. This process was performed every 10 frames, and the locations were remembered in the intermediate frames.

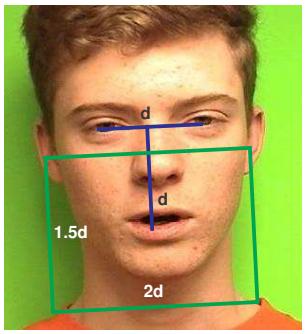

Fig. 5. Calculating lip search region from eye locations.

2.1.3. *Lip Location and Tracking*

Once the eye locations have been found, they are used to calculate a lip search region, as shown in Figure 5. The lip search region is then rotation-normalised, converted to R/G colour-space, and thresholded. The lip candidates from the thresholding are examined to remove unlikely lip locations (eg. too small, wrong shape). A search-window of 125×75 pixels is then scanned over the lip candidate image to find the windows with the highest concentration of lip candidate regions. The final lip ROI is chosen as the lowest, most central of these windows. Once the ROI was correctly located, it was rescaled to 60×36 pixels for the experiments.

2.2. Visual Speech Feature Extraction

Once each ROI was found for each frame of visual speech, features were extracted from it. Current area-based visual speech feature extraction techniques such as principal component analysis (PCA), linear discriminant analysis (LDA) and the discrete cosine transform (DCT) are all based on *monolithic* representations [13] of the mouth. The term monolith is used to describe the holistic vectorised representation of the mouth based purely on pixel values within an image array. The DCT is very widely used because, unlike PCA and LDA, is amendable to fast computations and avoids expensive training [14]. The DCT also obtains the best performance as was shown in [15]. For this reason, the DCT was chosen as the monolithic technique used for this experiment. In this paper, the 2D-DCT was performed

on each mean-removed mouth image. The mean-removed technique is a form of feature normalisation and helps obtaining vital dynamic speech features which is beneficial for speech recognition and helps remove unwanted redundant speaker information [15]. The top 20 DCT coefficients according to a zig-zag scan were retained as well as 20 delta features.

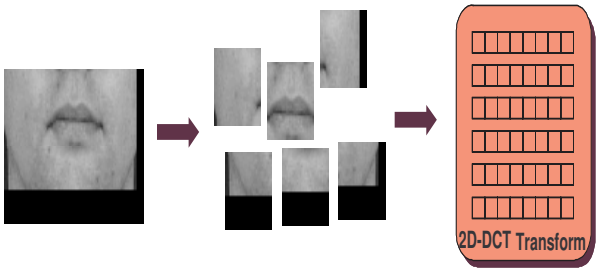

Fig. 6. Graphical depiction of the parts and feature representations of a face. Note: even though overlapping blocks are not depicted in practice the overlapping blocks leads to greater performance.

For the free-parts representation, each 60×36 ROI image was broken up into a series of 12×12 image patches with an overlap of 50%. The overlap between patches aids the process in two ways. Firstly, the overlap reduces the spatial area used to derive one feature vector and adds some redundancy between patches (i.e. no single patch contains all the information about a local region of the face). Secondly, as the overlap is increased it also increases the number of image patches (i.e. observations) exponentially. Once the image patches are acquired they then have an 2D-DCT applied to compact the $12 \times 12 = 144$ element patch into a feature vector $\mathbf{o}$ of suitable dimensionality to model a generalised but distinct distribution of that subject's face. The first 20 energy preserving 2D-DCT coefficients were extracted according to the zig-zag pattern. The differences of these patches over adjacent frames were also taken to give the deltas. This resulted each feature vector being size 40. A depiction of the feature extraction process for the free-parts representation is shown in Figure 6.

2.3. Visual Speech Modelling

In these experiments, each of the digits for the monolithic representation were modeled using a 9 state, left-to-right Hidden Markov models (HMMs) with 18 mixtures per state, using HTK [16]. This HMM topology was used as empirical and heuristic evidence showed that this gave optimal results. As mentioned previously, the 36 speakers were divided equally into a training set and a test set to provide a completely speaker-independent experiment. The HMMs were trained using this training set.

Modeling the free-parts representation was not so straightforward. This is because a modified form of the Viterbi and EM (expectation maximization) algorithms to enable the estimation of Free-Parts HMMs (FP-HMMs) had to be formulated. The modifications were necessary due to the inherent characteristics of a free-parts representation, namely: (i) That multiple observations will occur at the exact same time instant. (ii) A state transition within a FP-HMM can only occur when *all* observations for that mouth image have been evaluated (i.e. a state transition cannot occur given that only half the observations within a mouth image have been evaluated). The difference in topologies of a monolithic HMM and a FP-HMM is given in Figure 7.

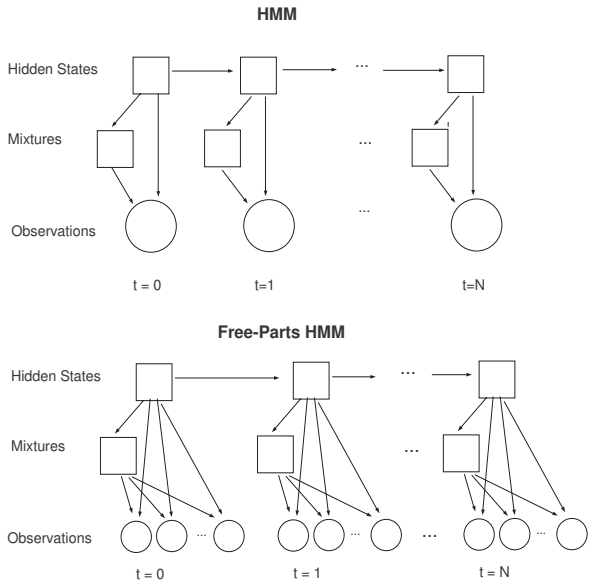

Fig. 7. Comparison of model topologies between a monolithic HMM and a FP-HMM.

Heuristically, we found the easiest way to accommodate these modifications was by resampling the images according to the number of free-parts patches per image. By performing the modeling in this fashion, the FP-HMMs could be modeled in a nearly identical way to the monolithic HMMs, which is important for comparative purposes. For example, in these experiments the frame rate was 29.97fps (or sampling period 33366.7 μs), and there were 45 patches per frame (or image). So by dividing the number of patches into the sampling period ($33366.7/45 = 741.48\mu s$), we found the new sampling period which was used to essentially give the equivalent modeling procedure to that of the monolithic.

3. RESULTS

The comparison between the monolithic and the free-parts representations were done in two sections. As previously mentioned, the first section was to compare the visual speech recognition performances of both the monolithic and free-parts representations on accurately detected and tracked ROI's and the second section was to perform the comparison of the poorly detected and tracked ROI's (see Figure 1). The poorly detected and tracked ROI's were obtained by randomly shifting the accurately tracked ROI's around a rectangle of 12×7 pixels.

The speaker-independent visual speech recognition results of the accurately tracked ROI's are given in Table 1. The train and test sets of each feature type were evaluated in terms of recognition rates. The difference between the train and test recognition rates are very important as this gives an indication of how undertrained a specific speech recognition classifier is using a certain type of feature [17]. The train recognition rate is also very important as it gives a rough estimate of the lower Bayes error for that feature representation, with the test recognition rates giving an estimate of the upper Bayes error. Both train and test errors are essential to properly evaluate a feature set.

As can be seen in Table 1, the train/test disparity in performance for the accurately tracked ROI's for the monolithic representation is quite large; indicating that the visual HMM classifier is under-trained for that representation. Contrastingly, the free-parts representation's train/test disparity in performance is small, thus validating to some extent our hypothesis that a free-parts representation naturally lends itself to more generalized classifiers. We should note that the capacity (i.e. the ability of the classifier to act as a lookup table) of the classifier being trained (i.e. HMM) will play a role in the interpretation of these results. This issue is important because the monolithic test performance is better than the free-parts test performance under ideal tracking conditions. However, one could argue that the capacity of the visual classifiers for both representation is roughly equal as we are employing a HMM classifier for both representations with the same parametric form.

Table 1. Speaker-independent digit recognition rates of monolithic representations compared to free-parts representations for accurately tracked ROIs.

	Train	Test
Monolithic	87.1%	39.0%
Free-Parts	25.3%	24.6%

Table 2, shows the speaker-independent visual speech recognition results for the poor tracking. In this experiment we found that the free-parts test performance slightly outperformed the monolithic test performance. This result suggests that our hypothesis of relaxing the spatial structure in the ROI to improve robustness against mistracking was correct. However, what is more interesting about these results is how much the performance of the monolithic suffers due to poor tracking. However, as mentioned in the introduction, this result was also predicted.

Table 2. Speaker-independent digit recognition rates of monolithic representations compared to free-parts representations for poorly tracked ROIs.

	Test
Monolithic	17.8%
Free-Parts	18.2%

4. CONCLUSIONS AND FURTHER WORK

In this paper, we have investigated the use of a free-parts representation for the task of speaker-independent visual speech recognition in an aid to alleviate some of the problems caused by the dependency that current monolithic representations have on accurate ROI detection and tracking. From the results, we have shown that a free-parts representation can marginally improve the recognition rates only in the extreme case where the ROIs were artificially detected and tracked poorly. However, it must be said that the overall performance of the free-parts representation strongly suggests that this type of representation does not preserve adequate speech information from the visual domain to perform speech recognition reliably. The loss of structure within the speaker's ROI, appears to have come at a cost with the loss the important speech information. This suggests that for visual speech recognition, structure is very important and may contain most of the important speech information. In contrast, as the monolithic representation is full of structure, constraining the full structure degrades performance greatly in scenarios where the detecting and tracking of the speaker ROI is inaccurate.

The results found in this paper are probably indicative of what the current level visual feature extraction is at the moment. To date, no representation of visual speech has shown itself to be speaker, pose, camera, and environment independent. Finding suitable representations which somehow encapsulate all these criteria will be a major focus of our future research. Maybe a potential solution could come in the form of some compromise between the free-parts and monolithic representations, where partial relaxation of the structure is relaxed to counteract the detecting and tracking inaccuracies whilst maintaining the vital speech information. However, as area-based visual representations used such as the monolithic and free-parts methods, closely parallels work done

previously in face recognition, it is of no surprise that these methods are highly speaker dependent and as such could be most useful in the field of audio-visual speaker recognition. Using these representations for the task of speaker recognition will be a future avenue of our research.

5. ACKNOWLEDGEMENTS

We would like to thank Clemson University for freely supplying us their CUAVE audio-visual database for our research.

6. REFERENCES

[1] F. Lavagetto, "Converting speech into lip movements: A multimedia telephone for hard hearing people," *IEEE Trans. Rehabilitation Engineering*, vol. 3, no. 1, pp. 90–102, March 1995.

[2] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," in *Issues in Visual and Audio-Visual Speech Processing,*. MIT Press, 2004.

[3] J. Luettin, N. A. Thacker, and S. W. Beet, "Speechreading using shape and intensity information," in *International Conference on Spoken Language Processing*, 1996, vol. 1, pp. 58–61.

[4] I. Matthews, T. Cootes, S. Cox, R. Harvey, and J. A. Bangham, "Lipreading using shape, shading and scale," in *Auditory-Visual Speech Processing*, Sydney, Australia, 1998, pp. 73–78.

[5] T. Wark and S. Sridharan, "An approach to statistical lip modelling for speaker identification via chromatic feature extraction," in *International Conference on Pattern Recognition*, 1998, vol. 1, pp. 123–125.

[6] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," in *Issues in Visual and Audio-Visual Speech Processing*, G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, Eds. MIT Press, Boston, 2004.

[7] G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," in *International Conference on Image Processing*, 1998, vol. 3, pp. 173–177.

[8] S. Lucey and T. Chen, "A GMM parts based face representation for improved verification through relevance adaptation," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[9] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.

[10] Z. Tufekci E. K Patterson, S. Gurbuz and J. N. Gowdy, "Cuave: A new audio-visual database for multimodal humancomputer interface research," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2002.

[11] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, "Moving-talker, speaker-independent feature study, and baseline results using the cuave multimodal speech corpus," *EURASIP Journal on Applied Signal Processing*, vol. 11, pp. 1189–1201, 2002.

[12] D. Butler, C. McCool, M. McKay, S. Lowther, V. Chandran, and S. Sridharan, "Robust face localisation using motion, colour and fusion," in *Seventh International Conference on Digital Image Computing: Techniques and Applications*, C. Sun, H. Talbot, S. Ourselin, and T. Adriaansen, Eds., Macquarie University, Sydney, Australia, 2003, CSIRO Publishing.

[13] S. Lucey, "The symbiotic relationship of parts and monolithic face representations in verification," in *International Workshop on Face Processing in Video (FPIV)*, Washington D.C., USA, 2004.

[14] G. Potamianos and P. Scanlon, "Exploiting lower face symmetry in appearance-based automatic speechreading," in *Proceedings of the Auditory-Visual Speech Processing International Conference 2005*.

[15] P. Lucey, D. Dean, and S. Sridharan, "Problems associated with current area-based visual speech feature extraction techniques," in *Proceedings of the Auditory-Visual Speech Processing International Conference 2005*.

[16] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book (for HTK version 2.2)*, Entropic Ltd., 1999.

[17] S. Cox, I. Matthews, and J. A. Bangham, "Combining noise compensation with visual information in speech recognition," *Auditory-Visual Speech Processing*, 1997.