Macro-cuboïd based probabilistic matching for lip-reading digits

Samuel Pachoud Shaogang Gong Andrea Cavallaro
Queen Mary, University of London
London, United Kingdom
{spachoud,sgg}@dcs.qmul.ac.uk andrea.cavallaro@elec.qmul.ac.uk

Abstract

In this paper, we present a spatio-temporal feature representation and a probabilistic matching function to recognise lip movements from pronounced digits. Our model (1) automatically selects spatio-temporal features extracted from 10 digit model templates and (2) matches them with probe video sequences. Spatio-temporal features embed lip movements from pronouncing digits and contain more discriminative information than spatial features alone. A model template for each digit is represented by a set of spatio-temporal features at multiple scales. A probabilistic sequence matching function automatically segments a probe video sequence and matches the most likely sequence of digits recognised in the probe sequence. We demonstrate the proposed approach using the CUAVE [23] database and compare our representational scheme with three alternative methods, based on optical flow, intensity gradient and block matching, respectively. The evaluation shows that the proposed approach outperforms the others in recognition accuracy and is robust in coping with variations in probe sequences.

1. Introduction

For perceiving facial emotion and behaviour, humans combine the acoustic waveform (audio information) and the movements of the lips, tongue and other facial muscles (visual information) generated by a speaker. The McGurk effect [19] established this bi-modal speech perception by showing that, when conflicting audio and visual stimuli are presented to an individual, the latter may assimilate a new stimulus, different from the other two.

Such observations have motivated interest in developing systems for automatic recognition of visual speech. Research in this field aims to improve speech recognition systems by taking advantages of the visual modalities of a speaker in addition to the usual audio modalities. Nevertheless, the performance of automatic lip-reading systems, i.e. speech recognition systems using visual information alone, is far from satisfactory. This underachievement is mostly due to the difficulty with finding a robust and consistent method to extract speech-relevant visual features.

Broadly speaking, there are four main approaches to performing lip-reading through visual recognition. The first one is grounded in texture-based visual features [10, 13, 20, 24, 25] assuming that all pixels encode visible speech data. With this approach, the features carry useful discrimination information and are estimated directly from a generally defined Region of Interest (ROI), such as the mouth, the lips, or the cheeks. Such methods then rely on traditional pattern recognition and image compression techniques (e.g. LDA, PCA, DCT, DWT) to reduce the high dimensionality and high redundancy of feature vectors and to extract relevant and useful lip-reading information. However, as all pixel values from the ROI are taken into account, it can contain irrelevant information. Moreover, texture-based systems are sensitive to intensity variations between the training and test data sets. On the other hand, methods based on shape visual features [1, 5, 12, 26] require adequate lip or mouth shape tracking, and assume that the visual speech information is captured by the form of the shape and the movement alone. Chen [5] and Kaynak *et al*. [12] use geometric features like outer-lip or inner-lip parameters. Alternatively, lip model features are used by Aleksic *et al*. [1] and Wang *et al*. [26], which consist of a model for the visible speech articulators, usually the lip contours. However, shape-based approaches suffer from complex feature extraction and training processes. Due to the nature of shape-based methodology, particular shapes adopted may not consider all relevant speech information and they are also over sensitive to image quality or resolution. Bridging these two extremes, various combinations of the two have also been used ranging from simple concatenation [5, 9] (combining two classes of features) to their joint modelling [18, 27]. The latter used an Active Appearance Model (AAM) [6], which provides a framework to combine shape and grey-level variation in a single statistical appearance model. Finally, motion based approaches [14, 17], which assume that visual motion during speech production contains relevant speech information, have been

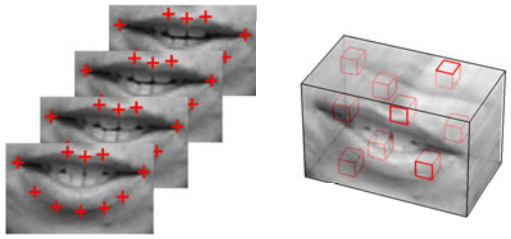
978-1-4244-2243-2/08/$25.00

proposed largely using optic flow. Current motion-based approaches mostly capture first order motion unable to cope with quick movements (e.g. certain parts of lip movement). A summary of lip-reading approaches is shown in Table 1.

Reference	Segmentation	Visual feature
		Texture
[24, 25]	DCT/DTW	LDA/MLLT
[20]	colour information	DCT / LDA
[8]	colour information	PCA / DCT
[13]	MESH / DFT	PCA / LDA
		Shape
[12]	-	height and width, area and angle
[5]	GMM	1 width and 2 height
[1]	-	Snake and parabolas
[26]	FCMS	ASM
		Texture and shape
[4]	GMM (colour space)	PCA
[18]	ASM/AAM/sieve	
[27]	AAM	
[9]	-	colour and geometric features
		Visual motion
[17]	-	optical flow
[14]	-	eigensequences

Table 1. A summary of lip-reading approaches divided into four main groups. AAM: Active appearance model; MESH: Collection of vertices and polygons; ASM: Active shape model; MLLT: Maximum likelihood data rotation; DCT: Discrete cosine transform; LDA: Linear discriminant analysis; DFT: Discrete Fourier transform; PCA: Principal component analysis.

In this paper we address the problem associated with most existing feature-based techniques, which assume continuous appearance of image patches over time in their entirety, and therefore tend to fail when the visual features are occluded or partially disappear. This failure is often the case in lip-reading due to deformation and self-occlusion. Moreover, most of the existing feature-based methods need manual labelling or alignment between frames. This is partially due to that given a shape-based model, one assumes the continuous existence of specific features over time such as the corners of the mouth or the outer-lip. We avoid such assumptions by adopting a set of built-up space-time volume features, which we refer to as macro-cuboïd (see Figure 1). The term macro-cuboïd comes as a spatio-temporal extension of macroblock[1], which is a widely used term in video compression. The proposed representation accommodates an arbitrary number of features in a given macro-cuboïd corresponding to different image regions of a moving lip. Furthermore, we introduce a probabilistic sequence matching function that allows for non-rigid multi-scale matching between different video sequences. With this approach, we aim to build a set of model templates consisting of a database of all the visemes[2] of the studied language. These templates provide a representation at an atomic level to be both concise and generative, because for instance the English language is composed of only about fifteen visemes [22]. To start, we focus in this paper on building model templates for automatically segmenting and recognising 10 digits appearing randomly in continuous probe video sequences.



(a) (b)

Figure 1. Examples of spatial only and spatio-temporal feature based lip-reading. (a) Image-to-image approach: features (red crosses) are extracted in each frame and then matched between frames. (b) Space-time volume modelling: features (red oblongs) are defined in space and over time to embed lip movements.

The remaining of the paper is organised as follows. Section 2 details our method for visual feature extraction and video sequence matching. Experiments and evaluation are presented in Section 3. Conclusions are drawn in Section 4.

2. Sequence matching for lip-reading of digits

We aim to recognise spoken digits through lip-reading. The proposed model is built given a database of model digit templates, the atomic level representing the digits 0 to 9 separately, before matching those model templates with probe video sequences. In a probe sequence, the order and the number of pronounced digits are unknown. Our model consists of two major parts: (1) visual feature selection and extraction, and (2) video sequence matching. Our first step defines and extracts automatically sets of spatio-temporal features, which we refer to as macro-cuboïd (see definition above), without any manual labelling of feature points, alignment between frames, or scale normalisation in space. The macro-cuboïds are then divided into a set of cuboïds, covering at least some parts of the lip movement. These cuboïds are represented at multiple spatial scales. Then a kernel-based maximum likelihood matching function is utilised to find the best match of all the macro-cuboïd candidates in a probe sequence for a model template. Digit recognition is determined by a histogram computed with the highest probability of a model macro-cuboïd (i.e. the

[1] Macroblocks (16x16 pixels) are used for motion estimation and compensation in traditional video encoders (H.261, MPEG-1/2).

[2] A viseme is a basic unit of speech in the visual domain that corresponds to phoneme (which is the basic unit of speech in the acoustic domain).

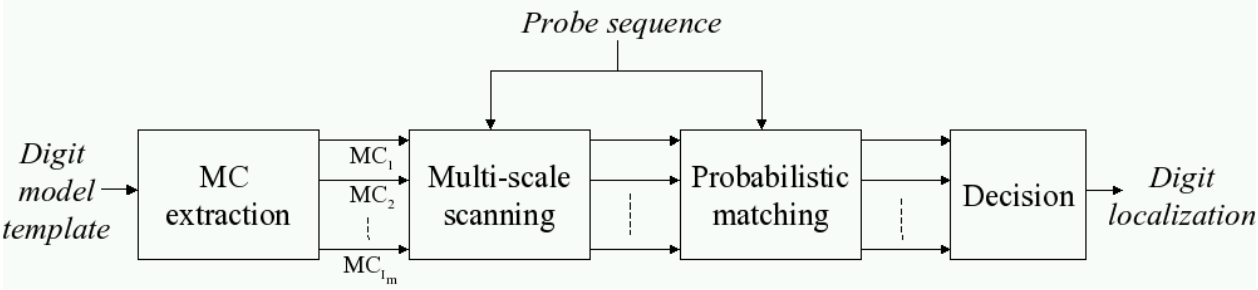


Figure 2. Processing blocks of our digit recognition system. Each digit model template is represented by a set of spatio-temporal macro-cuboïds (I_m macro-cuboïds). Then for each macro-cuboïd, we perform multi-scale scanning through the probe sequence. A probabilistic matching function is computed for each scan. Then a histogram-based decision is made to localise and recognise the model template.

biggest bin) indicating both the existence of a digit and its exact location in the probe sequence. Figure 2 gives an overview of our approach. We shall describe the details in the following.

2.1. Spatio-temporal approach

Instead of extracting the principal components of lip movement in order to establish a one-to-one correspondence between phonemes of speech and visemes of lip shape [11, 25, 26], here we consider comparing the movements of lips generated by a speaker (*probe* movements) with the movements of lips of particular words (digits in this paper) in a certain language (*model* movements). This consideration induces space-time features, which embed the lip movements. The idea of working in space and over time is exploited in [2, 3, 7, 21]. These approaches are based either on matching space-time trajectories of moving regions or on detection of interest points (features) within a stack of frames. This methodology contrasts with the matching of explicit landmark interest points (e.g. corners, edges), which is the basis of most feature-based image-to-image matching techniques [12, 18]. Figure 3 shows the ambiguity exhibited by using a landmark feature based approach as compared to that of a spatio-temporal model.

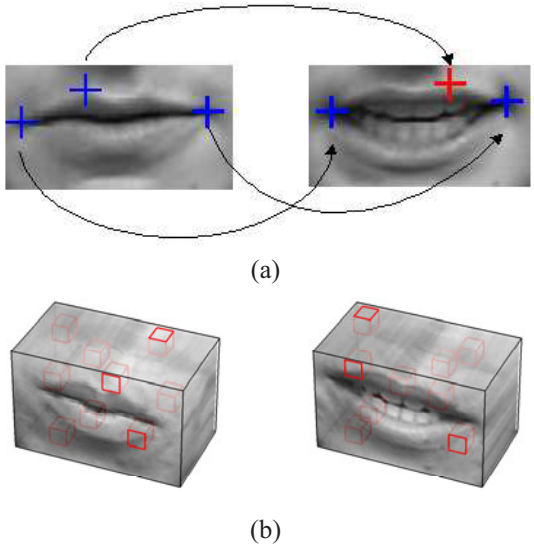


Figure 3. Spatial feature ambiguity compared with a spatio-temporal model.(a) Image-to image-approach: 2 separate frames of 3 moving features. (b) Space-time volume modelling: a sequence with several spatio-temporal features. When analysing only single frames, the information of the movements is lost and the matching between features is strenuous. On the other hand, space-time features are capable of capturing object movements.

2.2. Feature selection and extraction

A set of model digit templates is divided into several macro-cuboïds, which are automatically selected to cover the whole space and time of the model digit templates (exhaustive division). Matching a model template with a probe sequence requires the computation of a probability function between the extracted macro-cuboïds from each model within the probe sequence. For each model template, this operation is performed I_m times (see Figure 2), where I_m corresponds to the number of model macro-cuboïds over multiple scales of the m^{th} model template.

Considering a model template and a probe sequence, our algorithm works as follow: At first we divide the model template into I_m macro-cuboïds MC_i^m, with $i = 1, 2, \ldots, I_m$. The number of macro-cuboïds I_m is determined by the size, in space and in time, of the model template. Each model template can have a different value of I_m. Based on several experiments using different overlapping levels, we decided to use a fifty percent overlap in time and none in space between the macro-cuboïds in order to cover the entire model digit templates.

After the extraction, each macro-cuboïd is divided into N cuboïds, C_j^m, with $j = 1, 2, \ldots, J_m$. To be able to manage certain different sizes of the ROI, the cuboïds have multiple scales in space, Sc_x and Sc_y (currently two scales are used). The variations due to minor global movements

are coped by the macro-cuboïds rather than each individual cuboïd, therefore the cuboïds have a fixed scale in time, Sc_t (see Figure 4). The number J_m of cuboïds is determined by those three scales, Sc_x, Sc_y and Sc_t.

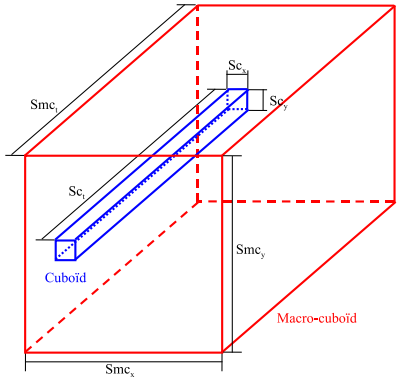


Figure 4. Macro-cuboïd MC_i^m and its respective cuboïd C_j^m (see Equation (2)). MC_i^m and C_j^m are an oblong but $Sc_t = Smc_t$. Two scales are used for Sc_x and Sc_y.

The initial scale value, S_1, is manually selected in the beginning of the process. Some preliminary experiments fixed the value of S_1. Then the second scale value is defined as follows: $S_2 = 2 \times S_1$.

2.3. Probabilistic sequence matching

After the selection and the extraction of the spatio-temporal features, we perform a probabilistic sequence matching. For each scale, S_1 and S_2, the probability of a model macro-cuboïds to be matched with the probe sequence, PS, is as follows:

$$P(\mathrm{MC}_i^m, \mathrm{PS}) = \frac{1}{\sigma_d^2 + \sigma_l^2} \sqrt[N]{\prod_i^N e^{-\frac{|\Delta d_i|^2}{2\sigma_d^2}} e^{-\frac{|\Delta l_i|^2}{2\sigma_l^2}}} \quad (1)$$

where Δd_i and Δl_i are respectively the differences and local displacements between descriptors of the cuboïds C_j^m and their correspondents in the probe sequence. σ_d and σ_l are the only parameters we have to define in Equation (1). σ_l is equal to the norm of the diagonal of macro-cuboïds MC_i^m. σ_d is determined empirically to give an equivalent weight of $-\frac{|\Delta d_i|^2}{2\sigma_d^2}$ with $-\frac{|\Delta l_i|^2}{2\sigma_d^2}$ in Equation (1).

Our cuboïd descriptors is an adaptation of Lowe's SIFT descriptor [15] for cuboïds [7]. At first we compute the gradient of cuboïds as follows:

$$G = \left[\frac{\partial \tilde{C}_j^m}{\delta x} \; \frac{\partial \tilde{C}_j^m}{\delta y} \; \frac{\partial \tilde{C}_j^m}{\delta t} \right] \quad (2)$$

where the first and the second derivatives are the differences in space, x and y and the last derivative is the differences in time t. Then the gradient G is divided into separate regions and a local spatio-temporal histogram is created for each region. The histograms values are then flattened into a vector to create a descriptor d_i for its corresponding cuboïd. The goal is to introduce robustness to small perturbations while retaining some positional information.

Finally, probability P_{tot} for matching a model macro-cuboïd to a segment of a probe sequence, taking into account two different scales of C_j^m, is:

$$P_{tot}(\mathrm{MC}_i^m, \mathrm{PS}) = max(P_{S_1}, P_{S_2}) \quad (3)$$

where P_{S_1} and P_{S_1} are the probabilities $P(\mathrm{MC}_i^m, \mathrm{PS})$ (Equation (1)) using the two scales of the cuboïd, S_1 and S_2, respectively. This probability P_{tot} is computed I_m times.

2.4. Digit recognition by lip-reading

At the end of the previous process, an histogram of the I_m model macro-cuboïds with the highest probability to be in the probe sequence is computed. The biggest bin indicates the position of the most likely match between a model digit template and a segment in a probe sequence. This information gives us the recognition and the localisation of digits in the probe sequence. If we assume that a set of model templates fully represents a language, then each part of a probe sequence can be decrypted. The model templates will consist of a database of all the visemes of the language. The main advantage of this is that the database will be concise and generative, because for instance, as mentioned earlier, the English language is composed of 15 visemes only [22]. For the examples used in this paper, we need to model 10 digits only in order to analyse any arbitrary combination of pronounced digits in a video sequence.

3. Experiments

3.1. Dataset

The number of existing audio-visual database is small compared to the number of audio-only speech databases, which have been collected for a longer time. For our experiments, we use the CUAVE database [23]. The CUAVE corpus is a moving-talker speaker-independent database, designed to support research into audio-visual speech recognition. The database consists of two major sections: one of individual speakers (36 different speakers) and one of speakers pairs (20 sequences). For both the individuals and groups sequences, connected and continuous digits between 0 and 9 are spoken while standing still in the first part of the clips. As this was not forced, there are some small, natural movements among these speakers. The last part of the clips is more challenging. In the individual speaker sequences, the speaker moves around intentionally while talking; in the multiple speaker sequences, two speakers are uttering the

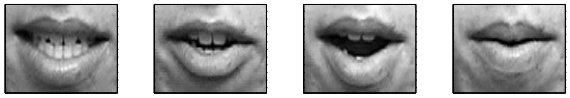
digits simultaneously. The individual sequences are about 2 minutes long and the group ones about 20-25 seconds long, at 29.97 fps (NTSC video standard).

3.2. Setup

The database is converted into grey-level images and each frame is cropped around the mouth (ROI). We divided the dataset into two parts: one part is used to generate the model templates and the other is used for the probe sequences. Each digit from 0 to 9 consists to one model template separately. To legitimate the fact that our method does not need any scale normalisation either in space or in time, we create several samples of each model digits. Hence each sample has a different size in space and in time (according to the pronunciation speed of the subjects). Figure 5 shows an example of frames for the model template representing the digit 0.

The probe sequences have a variable length, ranging from 4 to 10 digits in duration. The sorting of the digits can be either in an increasing order, in a decreasing order or at random. As for the model digit templates, each digit can have a different size in space and duration over time.

In the next two sub-sections, we present, at first, a comparative evaluation with three other methods and secondly we show different results with our approach, macro-cuboïds matching using spatio-temporal SIFT descriptor and local displacement, which we refer to as MCM-ST.



(a) (b) (c) (d)

Figure 5. Example of model template ROIs. (a) – (d): Samples from the normal model sequence for digit 0.

3.3. Comparative evaluation

We evaluate our approach by comparing it with three other different representations, two of which also macro-cuboïd based with the third one using a motion vector based strategy. The first alternative macro-cuboïd based representation calculates optical flow using Lucas & Kanade algorithm [16]. The second macro-cuboïd based method is based on histogrammed brightness values with twelve uniformly quantised bins. The last method uses the motion vector strategy. This spatial strategy consists of dividing an image into pixel macroblocks and then further dividing each macroblock into blocks. The sum of the differences between the pixel values of blocks between two consecutive frames is computed. The region with the smallest sum is chosen. We apply this strategy to our lip movements and we extend it to a spatio-temporal support. Instead of using blocks and macro-blocks we perform the motion vector strategy between the cuboïds of the model template and the probe sequence.

Figure 6 shows the confusion matrices for the four different methods. We can observe that our approach, MCM-ST, outperforms the others. For all the methods, the model digit 8 is not correctly localised. This error is due because the movements of the lip for the digit 8 in general are extremely limited. Consequently the matching is spread between every digits.

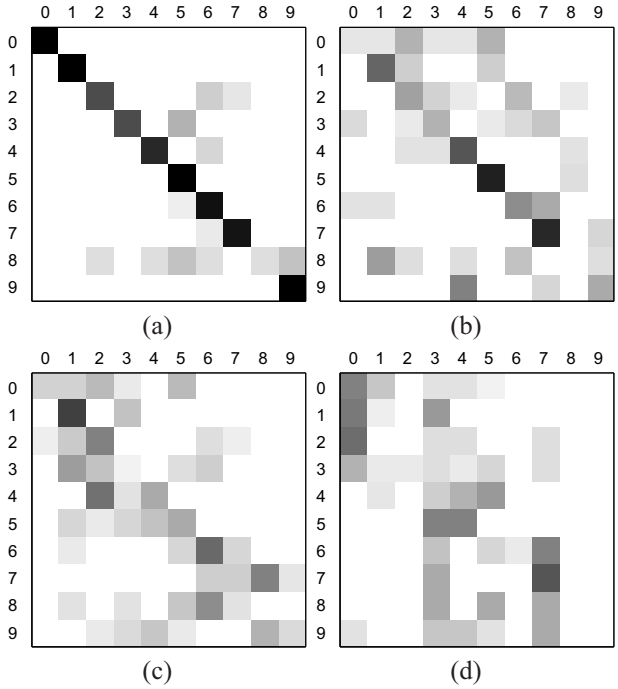


Figure 6. Confusion matrices from using (1) proposed method (MCM-ST), (b) optical flow, (c) intensity gradient, (d) motion vector matching. The columns represent the model template indices whilst the rows account for correctness of digit recognition and localisation.

3.4. Spatio-temporal space results

Figure 7 shows several examples of experimental results with nine different model templates on three different probe sequences. For those experiments, the speaker is male and the probe sequences contain 4 digits in duration. In each plot, we can see the probe sequence (shown by the biggest oblong), the macro-cuboïds with the highest probability to be in the probe sequence (the coloured oblongs) and the corresponding histogram. The biggest bin in each histogram indicates both the existence of the digit and its exact location in the probe sequence. In each probe sequence, the frame slices (4 slices per sequences) represent

the first frame of a digit. Therefore the representation of the experiments is more apparent.

We can observe in Figure 7 that the results are correct: all the digits, 3, 6 and 9, are localised correctly for each samples. The plots (c), (d) and (f), for instance, show the difference between the model template samples. Some samples for one digit are more representative of the digit. This observation could be used to generate one unique model template per digit (See Section 4). We can observe in plots (g) to (i) that when some digit (digit 2 and 3 in this case) are truly short in time (less than 7 frames), the macro-cuboïds are bigger than the digit itself. One solution is to reduce the scale in time of the macro-cubïds but this measure will remove too much useful information about the movements of lip in the features and increase the computational load. Other solutions need to be investigated.

Figure 8 shows similar results than Figure 7. The only difference is that the probe sequences contain 10 digits in duration. We observe that our approach, MCM-ST, is still able to recognise and localise the model digit even with different length of the probe sequences.

In both Figures 7 and 8, some macro-cuboïds are not correctly localised. Most of those macro-cuboïds are positioned in the edges of the ROIs, where only few movements exist. Adding a post-processing step to remove the macro-cuboïds on the borders would help limiting this issue.

Figure 9 shows a graph with the recognition rate for each model digit in the database. Each coloured curve represents another model digit. The x axe represents the number of macro-cuboïds used to recognise and localise a model digit in a probe sequence. The number 100% means we use all the I_m macro-cuboïds. We observe that if we decrease the number of macro-cuboïds used for the model digit localisation decision (x axe towards 0), the recognition rate decreases consistently.

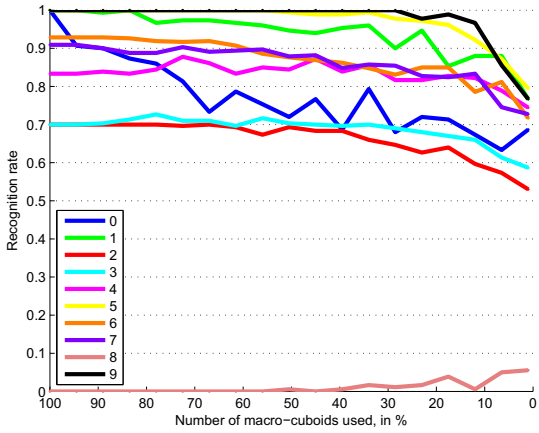


Figure 9. Recognition rates for matching single model digit templates representing all 10 digits with novel probe sequences of variable length, ranging from 4 to 10 digits in duration. The x axis shows the number of macro-cuboïds associated with the highest probability in matching model digits to probe sequences. The y axe gives the recognition rate, which also implicitly gives the most likely location of all the recognised digits in the probe sequences (see examples in Figures 7 and 8). Each curve represents a different model digit.

4. Conclusion

In this work we have shown the viability of a spatio-temporal feature-based lip-reading. These features embed lip movements from pronounced digits and contain more discriminative information than spatial features alone. This approach needs no manual labelling of feature points, no alignment between frames, nor scale normalisation in space, thanks to a macro-cuboïds based representation. This representation accommodates arbitrary number of features in a given cuboïds corresponding to different image regions of a moving lip.

Our system automatically selects spatio-temporal features extracted from 10 digit model templates and matches them into a probe video sequence. A model template for each digit is represented by a set of spatio-temporal macro-cuboïds at multiple scales. A probabilistic sequence matching function automatically segments a probe video sequence and matches the most likely sequence of digits recognised in the probe sequence.

The comparative evaluation between optical flow, intensity gradient, motion vector strategy and our method (macro-cuboïds matching using spatio-temporal SIFT descriptor and local displacement) shows that our method outperform the other approaches. Experimental results demonstrate that the existence of a model digit and its exact location can be found in a probe sequence.

Future works include an extension of the set of model templates to a database of all the visemes of a studied language. Also, the analysis of the results of the different samples of one particular digit needs to be investigated to be able to create one model template per digit.

References

[1] P. Aleksic, J. Williams, Z. Wu, and A. Katsaggelos. Audio-visual continuous speech recognition using MPEG-4 compliant visual features. In *ICIP*, 2002. 1, 2

[2] O. Boiman and M. Irani. Similarity by composition. In *NIPS*, 2006. 3

[3] Y. Caspi, D. Simakov, and M. Irani. Feature-based sequence-to-sequence matching. *IJCV*, 68(1):53–64, 2006. 3

[4] M. Chan. Hmm-based audio-visual speech recognition integrating geometric and appearance-based visual features. In *MSP*, 2001. 2

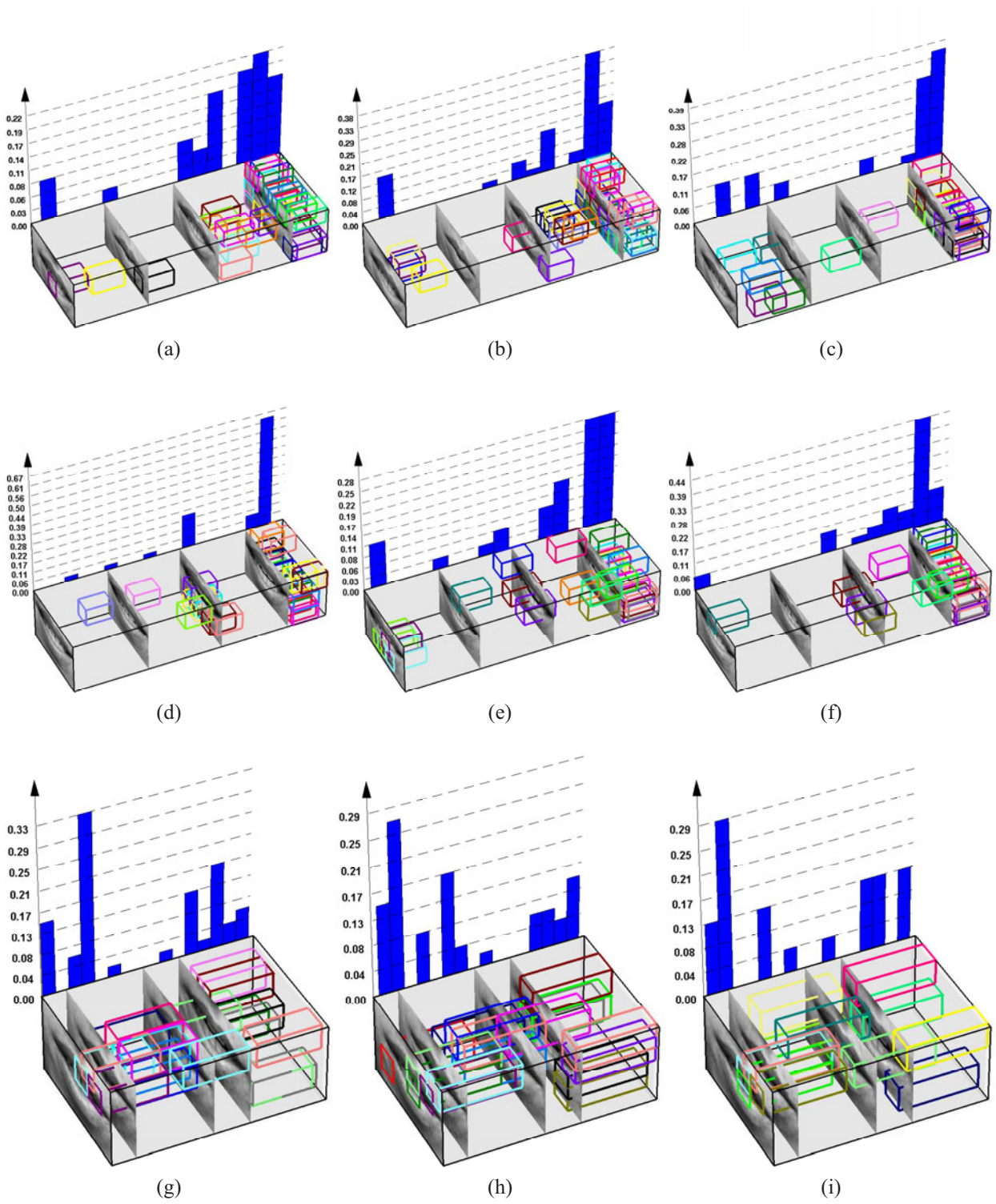
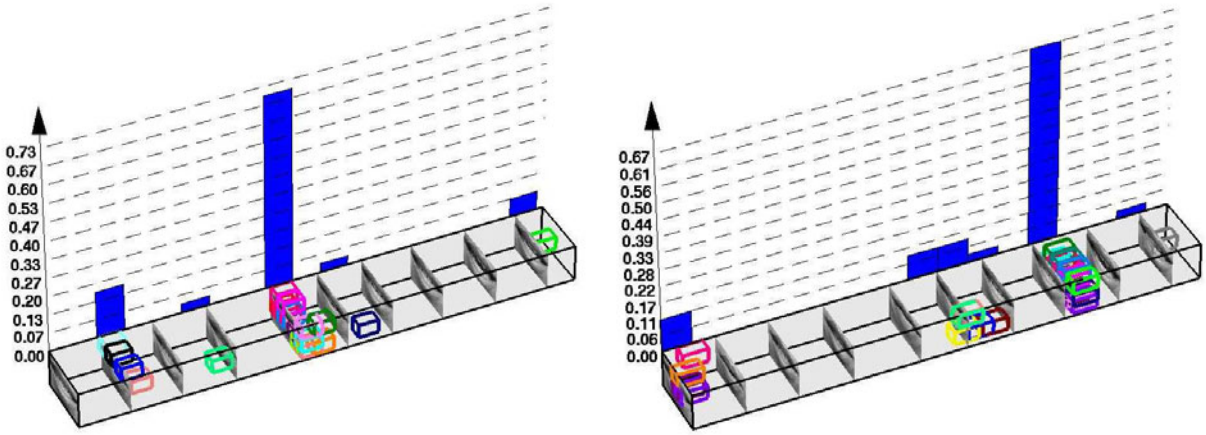


Figure 7. Experimental results using our approach, macro-cuboïds matching using spatio-temporal SIFT descriptor and local displacement. (a), (b) and (c) represent three different samples of the model digit 3 on the probe sequence with digits 0 , 1 , 2 and 3. (d), (e) and (f) represent three different samples of the model digit 9 on the probe sequence with digits 6 , 7 , 8 and 9. (g), (h) and (i) represent three different samples of the model digit 6 on the probe sequence with digits 2 , 6 , 3 and 7. In each plot, the biggest oblongs are the probe sequences (a frame slice represents the first frame of a digit), the coloured oblongs are the macro-cuboïds with the highest probability to be in the probe sequence and the graph is the corresponding histogram defined in Section 2.4.



(a) (b)

Figure 8. Experimental results using our approach, macro-cuboïds matching using spatio-temporal SIFT descriptor and local displacement. (a) represents one samples of the model digit 4 and (b) one sample of the model digit 7. Both of them are performed on the probe sequence with digits 0 to 9. In each plot, the biggest oblongs are the probe sequences (a frame slice represents the first frame of a digit), the coloured oblongs are the macro-cuboïds with the highest probability to be in the probe sequence and the graph is the corresponding histogram defined in Section 2.4.

[5] T. Chen. Audiovisual speech processing. *SPM*, 18(1):9–21, 2001. 1, 2

[6] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *PAMI*, 23(6):681–685, 2001. 1

[7] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VSPET*, 2005. 3, 4

[8] R. Göcke. Current trends in joint audio-video signal processing: a review. In *SSP*, 2005. 2

[9] R. Göcke. Audio-video automatic speech recognition: an example of improved performance through multimodal sensor input. In *NICTA*, 2006. 1, 2

[10] X. P. Hong, H. X. Yao, Q. H. Liu, and R. Chen. An information acquiring channel - lip movement. *ICACII*, 3784:232–238, 2005. 1

[11] J. Huang, G. Potamianos, and C. Neti. Improving audio-visual speech recognition with an infrared headset. In *ICAVSP*, 2003. 3

[12] M. Kaynak, Q. Zhi, A. Cheok, K. Sengupta, and K. C. Chung. Audio-visual modeling for bimodal speech recognition. In *ICSMC*, 2001. 1, 2, 3

[13] M. Leszczynski and W. Skarbek. Viseme recognition - a comparative study. In *AVSS*, 2005. 1, 2

[14] N. Li, S. Dettmer, and M. Shah. Lipreading using eigensequences. In *AFGR*, 1995. 1, 2

[15] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 4

[16] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, 1981. 5

[17] K. Mase and A. Pentland. Automatic lipreading by optical-flow analysis. *SC*, 22(6):67–76, 1991. 1, 2

[18] I. Matthews, T. Cootes, J. Bangham, S. Cox, and R. Harvey. Extraction of visual features for lipreading. *PAMI*, 24(2):198–213, 2002. 1, 2, 3

[19] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, December 1976. 1

[20] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy. Dynamic bayesian networks for audio-visual speech recognition. *EURASIP*, 2002(11):1274–1288, 2002. 1, 2

[21] J. C. Niebles, H. Wang, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. In *BMVC*, 2006. 3

[22] E. Owens and B. Blazek. Visemes observed by hearing-impaired and normal-hearing adult viewers. *JSHR*, 28(3):381–393, 1985. 2, 4

[23] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy. Cuave: a new audio-visual database for multimodal human-computer interface research. In *ICASSP*, 2002. 1, 4

[24] G. Potamianos, H. Graf, and E. Cosatto. An image transform approach for hmm based automatic lipreading. In *ICIP*, 1998. 1, 2

[25] G. Potamianos, C. Neti, G. Iyengar, A. W. Senior, and A. Verma. A cascade visual front end for speaker independent automatic speechreading. *IJST*, 4:193–208, 2001. 1, 2, 3

[26] S. Wang, W. Lau, S. Leung, and H. Yan. A real-time automatic lipreading system. In *ISCS*, 2004. 1, 2, 3

[27] W. Wang, D. Cosker, Y. Hicks, S. Saneit, and J. Chambers. Video assisted speech source separation. In *ICASSP*, 2005. 1, 2