Visual Speech Encoding based on Facial Landmark Registration

Ram P. Krish, Paul F. Whelan

Vision Systems Group, School of Electronic Engineering,
Dublin City University, Dublin, Ireland.
ram.krish@dcu.ie, paul.whelan@dcu.ie

Abstract

Visual Speech Recognition (VSR) related studies largely ignore the use of state of the art approaches in facial landmark localization, and are also deficit of robust visual features and its temporal encoding. In this work, we propose a visual speech temporal encoding by integrating state of the art fast and accurate facial landmark detection based on ensemble of regression trees learned using gradient boosting. The main contribution of this work is in proposing a fast and simple encoding of visual speech features derived from vertically symmetric point pairs (VeSPP) of facial landmarks corresponding to lip regions, and demonstrating their usefulness in temporal sequence comparisons using Dynamic Time Warping. VSR can be either speaker dependent (SD) or speaker independent (SI), and each of them poses different kind of challenges. In this work, we consider the SD scenario, and obtain 82.65% recognition accuracy on OuluVS database. Unlike recent research in VSR which makes use of auxiliary information such as audio, depth and color channels, our approach does not impose such constraints.

Keywords: Visual speech, temporal encoding, facial landmarks, dynamic time warping.

1 Introduction

Speech perception is the process by which the sounds of language is *heard, interpreted* and *understood*. The interpreting aspect also includes focusing on visual cues of the speech. The interactions between acoustic and visual information in speech perception was shown by McGurk, the phenomenon being popularly known as *McGurk Effect* [McGurk and MacDonald, 1976]. People with better sensory integration are more susceptible to McGurk effect. Visual cues generally used by humans for speech perception constitute *lip-motion, head movements, facial expressions, body gestures, language structures, contexts, etc.* Such a process is referred to as *speech reading* [Newman et al., 2010].

From an automated computational point of view in Visual Speech Recognition (VSR) where only visual cues are derived as features for recognition, lip-motion is considered more feasible compared to other visual cues. So, lip-motions encoded as visual features contributes towards VSR. When only acoustic

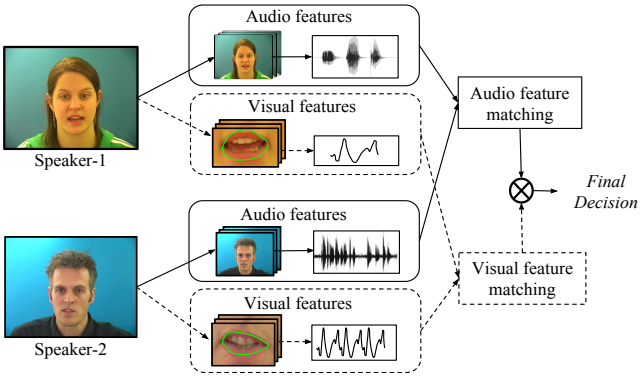

Figure 1: An Audio-Visual Speech Recognition (AVSR) system where the audio and visual (*lip-motion*) features of two speakers are compared to evaluate if they are similar or not. The images of speakers are taken from GRID audio-visual corpus [Cooke et al., 2006].

features are used in recognition, the system is referred to as Automatic Speech Recognition (ASR), and when both acoustic and visual features are used, the system is referred to as Audio-Visual Speech Recognition (AVSR). Figure 1 represents a general AVSR system comparing the audio-visual features of two speakers to evaluate the similarity in speech.

One of the major challenges in lip-motion analysis is due to *phonemes* and *visemes* not sharing one-to-one correspondence. A phoneme is one of the units of sound that distinguishes one word from another in a particular language. A viseme is defined as a visually distinguishable unit of speech in visual domain, the equivalent of phoneme in audio domain. Often, several phonemes correspond to single viseme [Cappelletta and Harte, 2012]. For example, words *pet, bell* and *men* are difficult to distinguish based on lip-motion because they have similar visemes while phonemes are different. The current *ARPAbet* phoneme set for standard English pronunciation maintained by Carnegie Mellon University Pronunciation Dictionary has 39 phonemes [Arp, CMU]. There is no standard viseme set similar to that of phonemes.

The detailed review on recent advances in the area of visual speech decoding [Zhou et al., 2014] points out that state of the art approaches to facial landmark localization is largely ignored in the development of VSR. The review also emphasized about the need for a better visual feature representation encoding the temporal information so as to improve the robustness of VSR. We address these two challenges in our work, and propose a methodology to improve VSR based on lip-motion analysis by incorporating a state of the art fast and accurate facial landmark detection which incorporates ensemble of regression trees learned using gradient boosting [Kazemi and Sullivan, 2014], as well as a simple temporal sequence encoding of visual features which can be verified using Dynamic Time Warping (DTW) algorithm.

The remaining part of this paper is organized as follows: a brief review of the related works in VSR, the OuluVS database used in our work, the proposed algorithm for temporal encoding of features corresponding to lip-motion (VeSPP), experiments demonstrating the robustness of the proposed method in speaker dependent scenario where the encoded features are compared using DTW, followed by conclusion and future work.

2 Related Works

Development of VSR in general involves visual feature extraction, its representation and classification. Extensive works was done on speech recognition based on audio signal alone, or on integrating audio and visual signals. Very little work has been reported in the literature for VSR alone. Various models for lip-motion analysis were studied by involving techniques such as Principal Component Analysis (PCA), Discrete Cosine Transform (DCT), Active Appearance Model (AAM), Hidden Markov Model (HMM), Local Binary Patterns (LBP), Support Vector Machines (SVM), etc. A detailed review of these techniques applied for lip-motion analysis can be found in [Zhou et al., 2014].

Of all the models, HMM is the most widely used technique in the domain of VSR [Liu and Cheung, 2014, Yu et al., 2009]. This is mainly due to the fact that HMM can incorporate strong temporal correlations between observed frames. However, the main challenges faced by HMM based VSR systems are: 1) the visual features obtained are not discriminant enough for lip-motion analysis and similarity computation, 2) the learned models are not sufficient to discriminate and characterize different lip-motion activities [Liu and Cheung, 2014].

In acoustic speech domain, there are well established features (for example, Mel-frequency cepstral coefficients (MFCC)), but in VSR, there are no standard accepted visual features. In general, visual features are broadly classified as *image-based, motion-based, geometric-based* and *model-based* [Zhou et al., 2014]. Many VSR based works in recent literature use auxiliary information such as audio corresponding to frames for preprocessing [Zhou et al., 2011], depth and color channel information [Pei et al., 2013] to accompany visual data. Though the systems which use such extra information report improved recognition accuracy, they tend to be more restricted in VSR domain, and these methods cannot be generalized. Also, usually in speaker dependent (SD) scenario, the number of training data available will be less than that of speaker independent (SI) scenario. This scarcity in training data is a major challenge for SD scenario.

In this work, we focus on the SD scenario and represent visual features using geometric-based attributes derived from facial landmarks corresponding to the lip region. These landmarks are obtained using an ensemble

of regression trees learned using gradient boosting as explained in [Kazemi and Sullivan, 2014]. The derived geometric features as well as their temporal encoding is described in the algorithm section, which is the main contribution of this paper. To these geometric features, we apply DTW to obtain a similarity score between any two given lip-motions. Our approach do not need any auxiliary information such as audio, depth or color channels. A similar purely visual only study was proposed by Zhao et al., where spatiotemporal local texture descriptors (LBP-TOP) are used for VSR [Zhao et al., 2009]. We will be following the experimental protocol and compare our results in SD scenario to the results reported in [Zhao et al., 2009].

3 Database

Although there are abundant audio-only databases for ASR, there exist only a few databases suitable for visual-only or audio-visual research. Among the audio-visual databases, many of them contain only recording of one subject, or are limited to isolated digits, letters or short list of fixed phrases not suitable for our experiments [Zhou et al., 2014]. There are few databases providing phrase data, but in many of them either the number of speakers is small or the speakers utter different phrases. For example, in GRID database [Cooke et al., 2006], all the phrases are different.

In this work we chose the OuluVS database which is publicly available and is a benchmark database in visual speech domain [Zhao et al., 2009]. It is a database containing the video and audio data for 20 subjects uttering 10 daily-use short phrases repeated up to 5 times making it suitable for visual speech lip reading experiments. The 10 phrases contained in OuluVS are: *Hello, Excuse me, I am sorry, Thank you, Good bye, See you, Nice to meet you, You are welcome, How are you, Have a good time*. The speakers were from 4 different countries with different accents and speaking rates which makes the dataset challenging. The videos were recorded in an indoor controlled environment. The frame rate was set as 25 fps and the image resolution was 720×576 pixels.

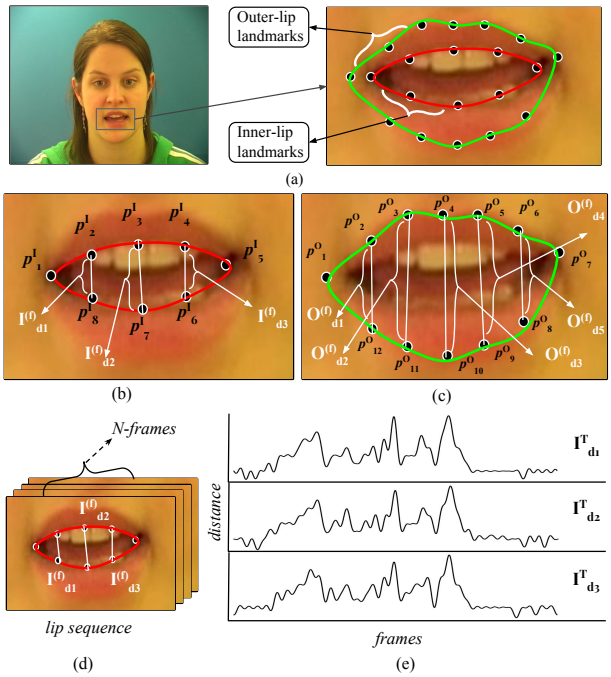

Figure 2: Various stages of the algorithm. (a) the facial landmark obtained for both inner and outer lip regions, (b) three vertical distances derived from the inner-lip landmarks, (c) five vertical distances derived from the outer-lip landmarks, (d) lip sequence consisting of N-frames, (e) temporal encoding of three vertical distances derived from the inner-lip landmarks for the video sequence consisting of N frames. The image of speaker is taken from GRID audio-visual corpus [Cooke et al., 2006] for demonstration purpose only.

4 Algorithm

The algorithm consists of four major stages: lip region landmark detection, visual feature extraction, visual feature encoding, and temporal sequence matching based on DTW for verification. Some of the recent work in facial landmark detection and lip-motion analysis can be found in [Kazemi and Sullivan, 2014, Katina et al., 2015, Sukno et al., 2015, Cao et al., 2014, Liu et al., 2015]. In this work, we used the algorithm proposed in [Kazemi and Sullivan, 2014] for facial landmark detection which used an ensemble of regression trees learned using gradient boosting. The Dlib C++ library [King, 2009] was used to train and obtain these landmarks for lip regions. Face detection was performed using Histogram of Oriented Gradients (HOG) as implemented in [King, 2009].

Stage 1: Landmark detection

Step 1: Using the algorithm proposed in [Kazemi and Sullivan, 2014], estimate the landmarks corresponding to the face.

Step 2: We detect 8 and 12 landmark points corresponding to inner-lip region and outer-lip region respectively (see Figure 2(a)). Depending on how the system is trained to obtain the landmarks, the number of points representing the lip region may vary. Let $p_i^I = (x_i, y_i)$ and $p_j^O = (x_j, y_j)$ represent the i^{th} and j^{th} landmark points for inner-lip and outer-lip regions respectively.

Stage 2: Visual feature extraction

Step 3: Locate the vertically symmetric landmark point pairs corresponding to inner-lip and outer-lip regions. In Figure 2(b), the vertically symmetric point pairs for inner-lip regions are (p_2^I, p_8^I), (p_3^I, p_7^I) and (p_4^I, p_6^I).

Similarly, for outer-lip region, the vertically symmetric point pairs are (p_2^O, p_{12}^O), (p_3^O, p_{11}^O), (p_4^O, p_{10}^O), (p_5^O, p_9^O), and (p_6^O, p_8^O).

Step 4: The distance $dist(p_k, p_l)$ between the vertically symmetric points (VeSPP) p_k and p_l represents a visual feature for a given frame. The vertical distance can be the *euclidean-distance* or the *absolute difference* between the symmetric points. In Figure 2(b), $I_{d1}^{(f)} = dist(p_2^I, p_8^I)$, $I_{d2}^{(f)} = dist(p_3^I, p_7^I)$ and $I_{d3}^{(f)} = dist(p_4^I, p_6^I)$ are the VeSPP features representing the inner-lip region for a given frame f.

Similarly, $O_{d1}^{(f)}$, $O_{d2}^{(f)}$, $O_{d3}^{(f)}$, $O_{d4}^{(f)}$, $O_{d5}^{(f)}$ are estimated from vertically symmetric points for the outer-lip region, as shown in Figure 2(c).

Stage 3: Visual feature encoding

Step 5: Assuming the lip sequence consists of N-frames (Figure 2(d)), repeat Steps $1-4$ for each frame and temporally concatenate the vertical distances corresponding to each vertically symmetric pairs to obtain its VeSPP temporal encoding. For example, $I_{d1}^T = \langle I_{d1}^{(1)}, I_{d1}^{(2)}, ..., I_{d1}^{(N)} \rangle$ represents the VeSPP temporal encoding for the vertically symmetric pair (p_2^I, p_8^I) corresponding to the inner-lip region for the lip sequence consisting of N-frames. Similarly, obtain the VeSPP feature temporal encoding of all vertically symmetric points corresponding to inner and outer lip regions.

Figure 2(e) shows the plot of the temporal encodings of I_{d1}^T, I_{d2}^T and I_{d3}^T for the inner-lip region thus obtained corresponding N frames (X-axis corresponds to frame number and Y-axis corresponds to vertical distance).

Stage 4: Similarity computation for verification

Step 6: Lip-motions representing a particular phrase by the same person at different instances may vary in both time and speed. The VeSPP features proposed in the previous steps encode the lip-motion as a temporal sequence. DTW can be used to find an optimal match between any two such encoded lip-motions. DTW is a dynamic programming based distance measure which allows a non-linear mapping of one temporal sequence onto another by minimizing the distance between them.

Suppose qI_{d1}^T and rI_{d1}^T represents the VeSPP temporal encoding of vertically symmetrical pairs (p_2^I, p_8^I) of two instances of lip-sequence videos consisting of M and N frames respectively where,

$$qI_{d1}^T = \langle q_1, q_2, q_3, ..., q_i, ..., q_M \rangle, \tag{1}$$

$$rI_{d1}^T = \langle r_1, r_2, r_3, ..., r_j, ..., r_N \rangle \tag{2}$$

where $q_i = I_{d1}^{(i)} = dist(p_2^I, p_8^I)$ of the i^{th} frame of the query video, and $r_j = I_{d1}^{(j)} = dist(p_2^I, p_8^I)$ of the j^{th} frame of the reference video.

These VeSPP features may correspond to lip-motions of the same speaker or different speakers.

To perform a non-linear alignment between qI_{d1}^T and rI_{d1}^T using DTW, we construct an $M \times N$ matrix where the $(i,j)^{th}$ entry of the matrix corresponds to squared distance $d(q_i, r_j) = (q_i - r_j)^2$ which is the alignment between q_i and r_j. The best match between qI_{d1}^T and rI_{d1}^T is found by retrieving a path through this matrix that minimizes the total cumulative distance between them. Essentially, the optimal path is the path that minimizes the warping cost

$$DTW(qI_{d1}^T, rI_{d1}^T) = min\Big(\sum_{k=1}^{K} w_k\Big) \tag{3}$$

where w_k is the matrix element $(i,j)_k$ that also belongs to the k^{th} element of a warping path W, a contiguous set of matrix elements that represents an optimal mapping between qI_{d1}^T and rI_{d1}^T. The warping path can be found using the dynamic programming to evaluate the recurrence

$$C(i,j) = d(i,j) + min\left\{\begin{array}{l} C(i,j-1) \\ C(i-1,j) \\ C(i-1,j-1) \end{array}\right\} \tag{4}$$

where $d(i,j)$ is the distance calculated for the current cell, and $C(i,j)$ is the cumulative distance of $d(i,j)$ and the minimum cumulative distance from the three adjacent cells.

Step 7: Let $c = DTW(qI_{d1}^T, rI_{d1}^T)$ be the minimum warping cost obtained, which is a dissimilarity measure, i.e., the lower the warping cost, the lower their dissimilarity which implies both temporal sequences are similar. We can transform the dissimilarity score to a similarity score S by

$$S = exp(-c) \tag{5}$$

Now, S close to 0 implies the temporal sequence compared are dissimilar, and when S is close to 1, the temporal sequences are similar. The ideal case is when same copies of signals are compared, which leads to a DTW value of 0 which in turn upper bounds to a similarity value of 1. The similarity scores thus obtained are normalized in the range $[0,1]$. So, our results can be directly used for multi-modal score fusions in AVSR applications.

5 Experiments

5.1 Experiment protocol

We used OuluVS video database in our experiments. The details about this database is briefly described in Section 3. We tested our proposed methodology for building a speaker dependent lip reading system. For each of the 20 speakers, the leave-one-video-out cross validation was carried out, i.e., one video is used for testing as a query template, and the rest were used as reference template. Since each of the 20 speakers uttered 10 different phrases repeated at least 5 times, there should be in total $20 \times 10 \times 5$ test comparisons in the cross validation scenario. In OuluVS video database, three video files corresponding to the repetition of three different phrases are not available. So, in our experiments, we have 997 test comparisons in total. For each testing, there are at most 4 match (*genuine*) scores and 19×4 non-match (*impostor*) scores. We determined whether the given comparison is a match or non-match based on the maximum similarity score obtained in the comparisons. We report the overall results of the cross validation in terms of recognition rate, obtained using M/N (M is the total number of correctly recognized sequence and N is the total number of testing sequence). Together with the recognition rate, we also generate the confusion matrix to see the clustering ability of the proposed method.

Configuration *(label)*	VeSPP feature configuration *(raw feature ⊕ first derivative)*	Recognition rate *(in %)*
C1	$(I^T_{d2}) \oplus (I'^T_{d2})$	71.61
C2	$(I^T_{d2} + O^T_{d2}) \oplus (I'^T_{d2} + O'^T_{d2})$	80.14
C3	$(I^T_{d2} + O^T_{d2} + O^T_{d3}) \oplus (I'^T_{d2} + O'^T_{d2} + O'^T_{d3})$	82.24
C4	$(I^T_{d2} + O^T_{d2} + O^T_{d3} + O^T_{d4}) \oplus (I'^T_{d2} + O'^T_{d2} + O'^T_{d3} + O'^T_{d4})$	**82.65**

Table 1: Recognition accuracy of speaker dependent experiments on OuluVS database for different configurations of VeSPP features.

Phrase	Excuse	Goodbye	Hello	How	Nice	Seeyou	Sorry	Thank	Time	Welcome
Excuse	**85.9**	4.0	0.0	0.0	7.1	1.0	2.0	0.0	0.0	0.0
Goodbye	3.0	**81.0**	1.0	2.0	5.0	1.0	0.0	0.0	0.0	7.0
Hello	2.0	0.0	**67.0**	0.0	1.0	15.0	1.0	14.0	0.0	0.0
How	0.0	0.0	2.0	**86.0**	2.0	2.0	2.0	3.0	1.0	2.0
Nice	2.0	0.0	0.0	4.0	**91.9**	1.0	0.0	1.0	0.0	0.0
Seeyou	1.0	0.0	9.1	1.0	2.0	**72.7**	0.0	13.1	0.0	1.0
Sorry	1.0	0.0	0.0	3.0	3.0	0.0	**88.0**	1.0	3.0	1.0
Thank	0.0	0.0	8.0	1.0	1.0	15.0	0.0	**74.0**	0.0	1.0
Time	0.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	**91.0**	6.0
Welcome	0.0	2.0	1.0	2.0	5.0	0.0	0.0	1.0	0.0	**89.0**

Table 2: Confusion matrix showing the recognition accuracy in percentage for the cross validation of 10 phrases uttered by 20 speakers of OuluVS database for the configuration C4 in Table 1.

5.2 Speaker dependent system

We used various configurations of VeSPP features and their first derivatives for testing the performance of the proposed visual speech encoding. In this experiment, we highlight only those VeSPP features which lead to best performance. So, the features discussed in this experiment are $I^T_{d2}, O^T_{d2}, O^T_{d3}, O^T_{d4}$, and we discard discussions about other VeSPP features. The first derivative of these sequences are also used which will be denoted as $I'^T_{d2}, O'^T_{d2}, O'^T_{d3}, O'^T_{d4}$. The first derivative is obtained by taking the difference between two consecutive values. We also generated a concatenated version of the features. Such concatenations of visual features were studied previously in [Zhou et al., 2011] and has shown performance improvements. We denote the concatenate version of I^T_{d2}, O^T_{d2} as $(I^T_{d2} + O^T_{d2})$. When we talk about match score for concatenated version $(I^T_{d2} + O^T_{d2})$, it is obtained by $DTW(qI^T_{d2} + qO^T_{d2}, rI^T_{d2} + rO^T_{d2})$ as mentioned in Eq.(3). When raw features and first derivative features are combined to obtained the match score, we denote it as $(I^T_{d2}) \oplus (I'^T_{d2})$, and the final score is obtained by adding the individual scores: $DTW(qI^T_{d2} + rI^T_{d2}) + DTW(qI'^T_{d2} + rI'^T_{d2})$.

Table 1 lists the recognition accuracy for various configurations of the VeSPP features. We used the *euclidean-distance* for VeSPP features. We noticed that the concatenated version of four features (configuration C4 in Table 1) taken from inner and outer lip regions together with their first derivative achieves the best result of 82.65% recognition accuracy. We also notice that for configuration C2, the performance is only slightly lower than that of C4. So, the proposed encoding scheme can be utilized for a faster implementation with a very small trade-off between the speed and accuracy.

We also report in Table 2 the confusion matrix to understand the clustering ability of our proposed method. The confusion matrix is for the 10 phrases in the OuluVS database by 20 speakers for the configuration C4 in Table 1 for the leave-one-video-out cross validation. The number at the i^{th} row and j^{th} column gives the percentage of i^{th} phrase being classified as j^{th} by our method.

Method	Recognition rate
[Zhao et al., 2009] (Automatic)	64.20%
[Zhao et al., 2009] (Manual)	70.20%
VeSPP method (Configuration C4)	**82.65%**

Table 3: Comparison with other visual only recognition accuracy for speaker dependent results for OuluVS database.

In Table 3, we compare our result with that of [Zhao et al., 2009] which followed the same experimental protocol as ours. In [Zhao et al., 2009], they propose two different experiments where the mouth region is manually located as well as automatically detected. In our experiment, mouth region is automatically detected, and we obtained 82.65% recognition accuracy compared to 64.20% obtained in [Zhao et al., 2009]. Also, our method outperforms the manually processed method which achieved 70.20% recognition accuracy.

In [Zhou et al., 2011], they report a much better result for automatic and manual processing for speaker dependent scenario, but those results cannot be compared to ours because, they used audio information to locate speaking and non-speaking frames, and then removed the non-speaking frames from the training video. So, their experiment is not purely visual only scenario and makes the audio information necessary to generate improved results. In [Pei et al., 2013], they proposed a method which uses depth information and color channels in VSR experiments, and their protocols were different. So, we discarded comparing our results to [Zhou et al., 2011] and [Pei et al., 2013].

6 Runtime analysis

The runtime complexity for facial landmark detection using ensemble of regression trees is a constant $O(TKF)$ where T, K and F are number of strong regressors, number of weak regressors and depth of trees respectively [Kazemi and Sullivan, 2014]. Deriving visual features corresponding to vertically symmetric pairs is a constant time operation, and is just taking absolute difference which is of $O(1)$. Once we have the temporally encoded VeSPP features, the verification can be performed using a fast-DTW comparison which can be performed in $O(N)$ time complexity, where N is the length of the temporal sequences [Salvador and Chan, 2004]. So, the proposed lip-motion verification can be achieved in linear time complexity upon detecting the face.

7 Conclusion and Future Work

We have proposed a robust temporal encoding of visual features (VeSPP) for lip-motion sequences based on distance computed from vertically symmetric points corresponding to lip regions. We used state of the art facial landmarks detection, and demonstrated its usefulness in lip-motion based verification using DTW comparison on a challenging database where the phrases are of different accents and speaking rates. Our experiments justify that concatenation of VeSPP visual features corresponding to inner-lip and outer-lip region provide better recognition accuracy and obtained 82.65% recognition rate. The fact that the proposed VeSPP features can be compared using DTW demonstrates its robustness in terms of negating the need for any training unlike HMM where more data samples are needed to train its model. In many real-time VSR applications, we cannot always expect to acquire more training samples, especially in case of speaker dependent scenario. Also, our method does not mandate any auxiliary information such as audio, depth or color channels, and is feasible on visual-only 2D data. We will be extending this work to build a speaker independent system based on the visual feature encoding developed in this work, as well as testing the system in real-time scenario.

Acknowledgments

This research was supported by funding from the charity RESPECT and the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement no: PCOFUND-GA-2013-608728.

References

[Arp, CMU] (CMU). Carnegie Mellon University Pronunciation Dictionary. `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`.

[Cao et al., 2014] Cao, X., Wei, Y., Wen, F., and Sun, J. (2014). Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190.

[Cappelletta and Harte, 2012] Cappelletta, L. and Harte, N. (2012). Phoneme-to-viseme mapping for visual speech recognition. In *ICPRAM (2)*, pages 322–329.

[Cooke et al., 2006] Cooke, M., Barker, J., Cunningham, S., and Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424.

[Katina et al., 2015] Katina, S., McNeil, K., Ayoub, A., Guilfoyle, B., Khambay, B., Siebert, P., Sukno, F., Rojas, M., Vittert, L., Waddington, J., et al. (2015). The definitions of three-dimensional landmarks on the human face: an interdisciplinary view. *Journal of anatomy*.

[Kazemi and Sullivan, 2014] Kazemi, V. and Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1867–1874. IEEE.

[King, 2009] King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758.

[Liu et al., 2015] Liu, H., Zhang, X., and Wu, P. (2015). Regression based landmark estimation and multi-feature fusion for visual speech recognition. In *IEEE International Conference on Image Processing*, pages 808–812.

[Liu and Cheung, 2014] Liu, X. and Cheung, Y.-M. (2014). Learning Multi-Boosted HMMs for Lip-Password Based Speaker Verification. *Information Forensics and Security, IEEE Transactions on*, 9(2):233–246.

[McGurk and MacDonald, 1976] McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*.

[Newman et al., 2010] Newman, J. L., Theobald, B.-J., and Cox, S. J. (2010). Limitations of visual speech recognition. In *AVSP*, page 1.

[Pei et al., 2013] Pei, Y., Kim, T.-K., and Zha, H. (2013). Unsupervised random forest manifold alignment for lipreading. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 129–136.

[Salvador and Chan, 2004] Salvador, S. and Chan, P. (2004). Fastdtw: Toward accurate dynamic time warping in linear time and space. In *KDD workshop on mining temporal and sequential data*, pages 70–80.

[Sukno et al., 2015] Sukno, F. M., Waddington, J. L., and Whelan, P. F. (2015). 3D Facial Landmark Localization With Asymmetry Patterns and Shape Regression from Incomplete Local Features.

[Yu et al., 2009] Yu, D., Ghita, O., Sutherland, A., and Whelan, P. F. (2009). A Novel Visual Speech Representation and HMM Classification for Visual Speech Recognition. In *Advances in Image and Video Technology*, pages 398–409. Springer Berlin Heidelberg.

[Zhao et al., 2009] Zhao, G., Barnard, M., and Pietikainen, M. (2009). Lipreading with local spatiotemporal descriptors. *IEEE Transactions on Multimedia*, 11(7):1254–1265.

[Zhou et al., 2014] Zhou, Z., Zhao, G., Hong, X., and Pietikäinen, M. (2014). A review of recent advances in visual speech decoding. *Image and Vision Computing*, 32(9):590–605.

[Zhou et al., 2011] Zhou, Z., Zhao, G., and Pietikäinen, M. (2011). Towards a practical lipreading system. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 137–144. IEEE.