JMLR: Workshop and Conference Proceedings 17 (2011) 33–40 2nd Workshop on Applications of Pattern Analysis

Employing The Complete Face in AVSR to Recover from Facial Occlusions

Ben Hall BEN.HALL@CS.UCL.AC.UK, **John Shawe-Taylor** JST@CS.UCL.AC.UK
and **Alan Johnston** A.JOHNSTON@UCL.AC.UK
University College London, London

Editor: Tom Diethe, José L. Balcázar, John Shawe-Taylor, and Cristina Tîrnăucă

Abstract

Existing Audio-Visual Speech Recognition (AVSR) systems visually focus intensely on a small region of the face, centred on the immediate mouth area. This is poor design for a variety reasons in real world situations because any occlusion to this small area renders all visual advantage null and void. This is poor by design because it is well known that humans use the complete face to speechread. We demonstrate a new application of a novel visual algorithm, the Multi-Channel Gradient Model, the deploys information from the complete face to perform AVSR.

Our MCGM model performs near to the performance of Discrete Cosine Transforms in the case where a small region of interest around the lips, but in the case of an occluded face we can achieve results that match nearly 70% of the performance that DCTs can achieve on the DCT best case, lips centric approach.

Keywords: AVSR, Facial Motion,

1. Introduction

Over the last several decades audio speech recognition has progressed a great deal, and although incremental performance gains are being achieved for niche scenarios, performance has plateaued for the most successful applications. In an effort to improve this performance researchers have focussed on the bimodal nature of speech perception in human listeners, incorporating and combining both the acoustic signal information and visual motions of the face information (McGurk and MacDonald, 1976). The inclusion of this information has been shown over the last several decades to improve the performance of human speech perception (Campbell, 2008).

This inclusion of visual information in speech recognition has lead to the field of audio visual speech recogniser (AVSR) systems (Potamianos et al., 2003). This inclusion of extra visual information provides a source of orthogonal information (Luttin and Thacker, 1997) that improves the performance of ASR to levels closer to that of human speech perception.

2. Background

2.1. Bimodal Speech Perception

The most striking example of the bimodal nature of speech perception in humans is by demonstration of the McGurk effect(McGurk and MacDonald, 1976). The McGurk effect is observed from the perceptual integration of visible open mouth syllable with an acoustically similar syllable but not visual corresponding.

Technically this is most easily demonstrated when an acoustic syllable is dubbed over facial motions of an alternative syllable. The most widely repeated example of the McGurk effect is that of the dubbing of 'ba' over the acoustic 'ga' leads to the perceived perception of 'da'(Campbell, 2008).

2.2. Facial Information in Bimodal Speech Perception

In an analogous manner to these investigations, research efforts have been focused on emulating these techniques in the visual domain. Preminger *et al.* selectively masked aspects of the face during speech production. This was achieved by digitally altering the raw image data, to a non-descript grey level and as such all information is removed from the segment in the face.

Greenberg and Bode also investigated the importance of facial cues other than those directly exhibited by the lips and mouths. This was demonstrated by occluding the top portion of their face to consider the case where the whole face is visible and a diminished case where the only the lips, mandible and larynx were visible. Greenberg and Bode tested consonant recognition, with 32 fully hearing subjects employing a single videotaped speaker. They measured a small but statistically significant improvement in recognition when the whole face was visible.

2.3. Existing AVSR Methodology

Using observation of the bimodal nature of speech the first AVSR system was demonstrated by Petajan (Petajan, 1984) in 1984, where he improved the performance of a single-speaker, isolated word recognition task on a 100-word vocabulary that includes letters and digits. Petajan's work was both the first example of the employment of visual information in conjunction with ASRs and simultaneously the first example of a *shape based* feature extraction technique (Yuille et al., 1992) on the visual stream.

In creating and combining feature extraction, audio-visual fusion and classification this system Petajan was simultaneously defining the basic template and methodology of the vast majority of future AVSRs (Potamianos et al., 2003). In an effort to parameterise the visual sequence Petajan was also the first to crop images of the whole face down to an isolated mouth and chin region-of-interest (ROI) (Gurbuz et al., 2001).

His reasons for this are self evident, despite there patently being significant amounts of visual facial motion and information expressed in the excluded face regions (Rosenblum and Saldana, 1996).

2.4. Classification

A vast majority of speech classification approaches have been proposed for both ASR as well as AVSR, the list of attempted methods includes artificial neural networks (Bregler and Konig, 1994), Support Vector Machines (Gordan et al., 2002) and visual feature space distance metrics (Petajan, 1984), but by a significant margin the most widely used approach is that of Dynamic Bayesian Networks (BDN) and specifically the use of Hidden Markov Models (HMMs)(Rabiner, 1990).

HMMs model the statistical transition between the various speech states and fundamentally assume a class-dependent generative model for the observed features.

3. Vision Processing

Rosenblum's *et al.*(Rosenblum and Saldana, 1996) and Jordan and Sergeant (Jordan and Sergeant, 2000) work demonstrate that existing AVSRs vision processing algorithms do not optimally employ the visual information and disregard amounts significant information away from the immediate mouth area.

3.1. A Different Methodology

An improved visual processing framework could thus successfully capture the global face movements and be able to employ the motions of the face to provide additional information about the speech patterns being spoken. A not completely dissimilar issue is addressed by Berisha *et al.* in his thesis

(Berisha et al., 2006), where they consider the problem of accurately recovering the full facial motions from occluded talking faces for use in conjunction with facial mimicry applications.

3.2. Berisha's Hypothesis

In Berisha's work an occlusion is artificially (digitally) placed over the lip and mouth region of the face over the duration of the subject speaking. Using the motions in the unoccluded areas of the surrounding face and a single static example of the unoccluded face enables the complete motions of the driving face to be accurately recovered (Berisha et al., 2006).

Figure 1: Example of Berisha's work. In the left hand side frame we observe a face with an artificially placed occlusion. The left hand side frame is the original image data and the central frame is the restored face after Berisha's processing (Berisha et al., 2006).

To perform this Berisha processes the driving sequence of frames with the Multi-Channel Gradient Model (McGM) (McOwan and Johnston, 1995) to derive a series of optical flows between frames. after employing this optical flow information Principle Component Analysis (PCA) vectors (Pearson, 1901) and Independent Component Analysis (ICA) (Comon, 1994) vectors are derived to provide efficient parameterisation of the data.

The sequence of images are projected into these vector spaces, and the points are then used to create a driver (Berisha et al., 2006) that is then used in conjunction with a single static frame to create a novel sequence of images.

3.3. Multi-Channel Gradient Model

In Berisha's method the image sequence is processed using McGM to model the optical flow between any particular frame against an average face, which has been derived in preprocessing from the complete sequences of faces. This difference calculation accurately captures enough relevant global (whole) facial motion that this image sequence can be used to drive an avatar where the lip and mouth motions have been restored.

Because of this success in facial mimicry it seemed plausible that McGM could provide the required information global information to allow for reconstruction of the lip movements when used as the visual front end to a visual noise resistant AVSR system.

3.4. Foundations of McGM

McGM model is derived from the physiological and psychophysically research attempting to understand the vision pathway in the human brain (McOwan and Johnston, 1995). It is well known (Hubel and Wiesel, 1977) that spatial domain of the receptive fields in V1 cortical area of the human brain are accurately modelled by derivatives of Gaussian functions (Koenderink and van Doom, 1987). These

cells behave in a manner similar to that of a convolution of Gaussian functions and the input images as a spatio-derivatives.

As the order of derivative filters is increased the filters become increasing tuned to higher spatial frequencies, forming a range of independent spatial channels.

The relationship between these measured temporal frequency sensitivity functions has been modelled by the process of temporal differentiation, however, combining these observations of temporal and spatial channels has proved problematic for physiology.

$$I(x+dx,t+dt) = I(x,t) + \frac{\delta I(x,t)}{\delta x}dx + \frac{\delta I(x,t)}{\delta t}dt + O(dx^2, dt^2) \tag{1}$$

To combine these sets of velocities estimates a least squares fit is performed to find the maximum likelihood value of the velocity v'. To accomplish this two vectors X and T may be constructed. The vector X comprises ordered components ($D_x\ I$, $D_{xx}\ I$, ..., $Dn_x\ I$), first order spatial to n^{th} order spatial, where $D_x\ I$, is the partial derivative of I with respect to x. Similarly the vector T is given by ($D_t\ I$, $D_{tx}\ I$,..., $D_{t(n-1)x}\ I$), first order temporal, first order spatial to first order temporal, $(n-1)^{th}$ order spatial. The best approximation to velocity v' contracts these two vectors to minimise the metric distance $(v'X - T)$ which requires $v' = (XT/XX)$, the denoting a scalar product. Explicitly this equation is shown in (2).

$$v' = \frac{\sum_n \frac{\delta^n}{\delta x^n} I \frac{\delta^{n-1}}{\delta x^{n-1}} \frac{\delta}{\delta t} I}{\sum_n \frac{\delta^n}{\delta x^n} I \frac{\delta^n}{\delta x^n} I} \tag{2}$$

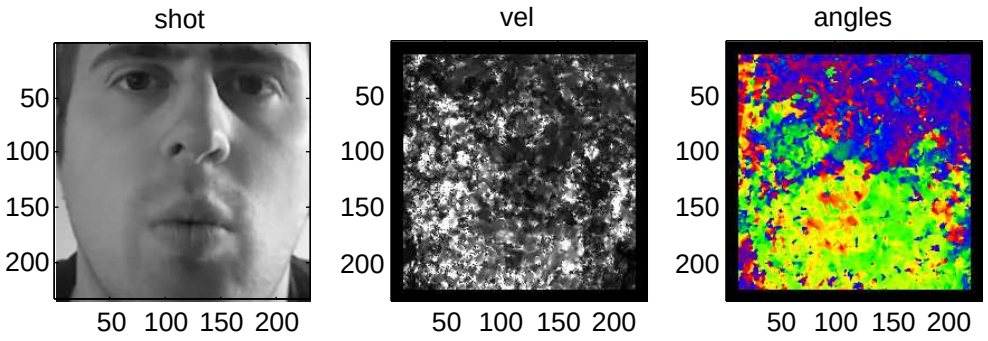

Figure 2: Example of the speed (middle frame) and angular (right handside frame) from computing MCGM. The left handside frame is the a superposition of the two frames that are being compared. In this example there is a small degree of movement on the lower jaw (N.B the large area of similar angular movement in the lower fraction of the right hand frame) and a slight protrusion of the lips that is vaguely visible approximately one third of the way up the same frame.

Calculation of this vector generates speed and angular information for every pixel over the image, this is demonstrated in Fig. 4.

4. Method

4.1. Vision Preprocessing

Before McGM can be employed on the face of speakers, the faces must be identified within the complete visual frame. To establish the facial region within the frame we have employed an implementation of the Viola/Jones algorithm (Viola and Jones, 2001), making use of OpenCV (ope).

Identification of the lips within the frame was via a custom dedicated Support Vector Machine (SVM) based lip detector (Chang and Lin, 2011).

4.2. Hidden Markov Model

Hidden Markov Models are trained for each digit in the database, using a left-to-right restriction on the transition matrix. HMMs are well understood and as such a full derivation is not given here.

4.3. Visual Features

For visual features the the visual input is passed through the MCGM model and mapped on to PCA vectors that are used as inputs to the HMM system.

4.4. Sound Parameterisation

The sound is parameterised, an extensively used form of audio parameterisation for Automatic Speech Recognition (ASR) are mel-frequency cepstrum coefficents (MFCCs) (Davis and Mermelstein, 1990). Mel-frequency cepstrum coefficients are calculated using time windows of 25ms with a 10ms hamming window overlap. MFCCs are a well established parameterisation of acoustic signals, so a full derivation is unnecessary.

4.5. Fusion Methods

To fuse between the acoustic and visual modalities we developed the two separate HMM streams and then seek to use late-stage decision fusion to determine the final classification.

5. Audio-Visual Database

As esteemed researchers have pointed out (Potamianos et al., 2003) it is very difficult to directly compare these methods due to the lack of common audio-visual databases.

5.1. HEBA Database

To correctly test our implementation of McGM, we required a database that included whole face information including many utterances from a speaker. To enable this we collected a corpus of full frontal video and audio of a single speaker uttering isolated digits ‘zero’, ‘one’, ‘two’, ‘three’, ‘four’, ‘five’, ‘six’, ‘seven’, ‘eight’, ‘nine’.

The database will be made available online (ben.hall[at]ucl.ac.uk, 2010).

6. Testing Possibilities

6.1. Occlusion Possibilities

The benefit of employment of McGM was the manner in which it has been successful used to reconstruct the occluded motions of the mouth using the unobscured movements of the remaining visible portion of the face.

6.2. Establishing a Baseline of Results

To establish a baseline for comparison to more advance techniques we consider the simplest case of classifying an audio only and combined audio-visual signal with a segmented ROI in the area immediately surrounding the mouth and lips region. To establish the visual components ability to clarify the audio modality, acoustic white noise is added to audio signal at varying SNRs

We compare the performance of the MCGM against that of the widely employed Discrete Cosine Transforms (type 2) (DCT-II). For the case of a ROI centered on the mouth and lips the performance of DCT is approximately 35% better than MCGM model, with DCTs seeing an approximate 10dB improvement compared to an equivalent 6dB improvement when McGM is used in conjunction with audio signal.

As well as providing a baseline comparison between DCT2 and McGM, these result provides some measure of the validity of the database since DCT-II has been extensively tested on the a variety of database and the results from this database and comparable to other similar database results.

6.3. Experimentation Set-Up

To test and examine the occlusion recover hypothesis we obscure the mouth and lip region and then recompute the optical flow, HMM framework and the associated speech classification.

There is the subjective factor in this set up: determining the segment of the face that the mouth and lip 'regions' consists of. In an effort to manage this subjectivity we have have used the lip bounding box, as one testing frame and everything else not included as the the other testing frame. This provides the guarantee that no facial information is repeatedly used in the different views. This provides a methodology of sorts for other researchers to compare against.

6.4. Results from Occlusion Methods

Employing this outlined methodology, we see a improvement in speech classification with the use of an occluded face, although the improvement is less than the improvement from a mouth centric ROI. As in the case of whole and face and mouth centric ROI the effect is most pronounced at lower signal to noise ratios.

6.5. Comparisons to DCT

Despite these obvious disadvantages DCT-II there is a small improvement when including the occluded visual information, but this is significantly less than McGM improvement.

Arguably the results seen from DCT and McGM, are as one would expect. Where DCT is selected for AVSR results because of it's ability to efficiently compress the variance of the source into a small number of vectors, where as McGM examines a broader area under the guise of temporal and spatial derivatives.

6.6. Commentary on Occlusion Methods

As to why these results are not as successful as the mouth centric ROI or even a full face ROI, despite there being 'enough' sufficient motion as demonstrated using Berisha's *et al.* work(Berisha et al., 2006). In Berisha's *et al.* work have an extra set of tuneable parameters which control the amount of facial deformation, via a scaling and application of PCA and ICA vectors to control the driven avatar. In doing this they can underplay or exaggerate (caricature) movements of the face, in the existing framework we do not have a directly analogous parameter and it seems plausible this has a noticeable effect.

7. Discussion

Although this methodology has shown somewhat the viability of employment of the complete face as the starting point for feature extraction, the conditions in which it has been deployed are very limited; the major limitations being a small vocabulary and the testing database being only on a single speaker. To fully test this idea we must expand both the vocabulary size and the amount of speakers.

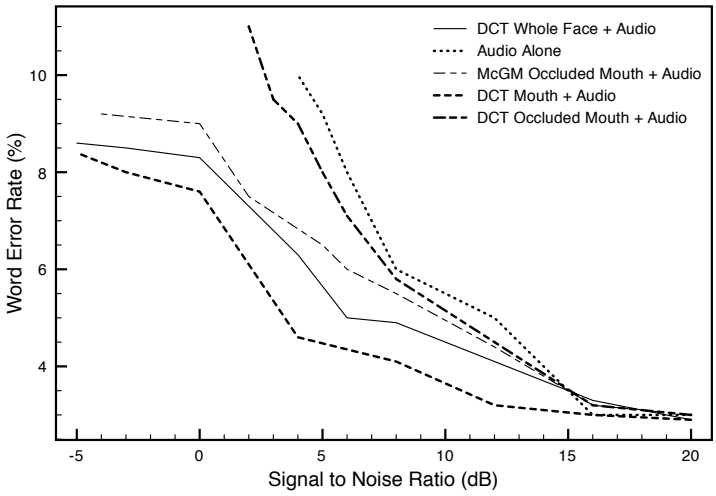

Figure 3: Audio-only and Audio-Visual ASR on the HEBA database. In the case of the AV ASR, a two stream HMM approach is employed, with decision based fusion used to determine the optimal classification combing both the audio and visual modalities.

This method has not been fully tested against more extensive challenges, but hopefully has provided some evidence that it is possible to include full face information can provide a degree of resistance to facial occlusion, which will have real world usages.

Acknowledgments

Thanks to CoMPLEX and ESPRC for funding

References

http://opencv.willowgarage.com/wiki/.

F. Berisha, A.Johnston, and P. McOwen. *Facial Mimicry*. PhD thesis, University College London, 2006.

Christoph Bregler and Yochai Konig. "eigenlips" for robust speech recognition, 1994.

R. Campbell. The processing of audio-visual speech: empirical and neural bases. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1493):1001–1010, 2008.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

P. Comon. Independent component analysis: a new concept? *Signal Processing*, 36:287–314, 1994.

Steven B. Davis and Paul Mermelstein. *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*, pages 65–74. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990. ISBN 1-55860-124-4. URL http://portal.acm.org/citation.cfm?id=108235.108239.

M. Gordan, C. Kotropoulos, and I. Pitas. A support vector machine-based dynamic network for visual speech recognition applications. 2002(11):1248, November 2002.

S. Gurbuz, Z. Tufekci, E. Patterson, and J. N. Gowdy. Application of affine-invariant fourier descriptors to lipreading for audio-visual speech recognition. In *in Proc. Int. Conf. Acoust., Speech, Signal Processing*, pages 177–180, 2001.

D. H. Hubel and T. N. Wiesel. Ferrier Lecture: Functional Architecture of Macaque Monkey Visual Cortex. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 198(1130):1–59, 1977. doi: 10.1098/rspb.1977.0085. URL http://rspb.royalsocietypublishing.org/content/198/1130/1.abstract.

T.R. Jordan and P. C. Sergeant. Effects on visual and audiovisual speech recognition. *Lang. Speech*, 43:107–124, 2000.

J J Koenderink and A J van Doom. Representation of local geometry in the visual system. *Biol. Cybern.*, 55(6):367–375, 1987. ISSN 0340-1200. doi: http://dx.doi.org/10.1007/BF00318371.

Jurgen Luttin and Neil A. Thacker. Speechreading using probabilistic methods, 1997.

Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264(588):746–748, 1976.

P.W. McOwan and A. Johnston. The algorithms of natural vision: the multi-channel gradient model. *IEE Conference Publications*, 1995(CP414):319–324, 1995. doi: 10.1049/cp:19951069. URL http://link.aip.org/link/abstract/IEECPS/v1995/iCP414/p319/s1.

K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 6: 559–572, 1901.

E.D. Petajan. Automatic lipreading to enhance speech recognition. In *Proceedings of the IEEE Communication Society Global Telecommunications Conference*, Atlanta, Georgia, November 26-29 1984.

G Potamianos, C. Neti, G. Gravier, A. Garg, and AW. Senior. Recent advances in the automatic recognition of audio-visual speech. *Proceedings of the IEEE*, 91(9), 2003.

Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. pages 267–296, 1990.

L.D. Rosenblum and H.M. Saldana. An audiovisual test of kinematic primitives for visual speech perception. *J. Exp. Psychol Hum. Percept. Perform*, 22:318–331, 1996.

ben.hall[at]ucl.ac.uk. Heba dataset, 2010.

P Viola and M Jones. Robust real-time object detection. In *Second Internation Workshop on Statistical and Computational Theories of Vison - Modeling, Learning, Computing and Sampling*, Vancouver Canada, July 2001.

A.L. Yuille, P.W Hallinan, and D.S. Cohen. Feature extraction from faces using deformable templates. *Int. J. Comput. Vision*, 8(2):99–111, 1992. ISSN 0920-5691. doi: http://dx.doi.org/10.1007/BF00127169.