SPEECH
COMMUNICATION
An international journal of the European Association for Signal Processing (EURASIP) and of the International Speech Communication Association (ISCA)

A Cascade Gray-Stereo Visual Feature Extraction Method For Visual and Audio-Visual Speech Recognition

Chao Sui, Roberto Togneri, Mohammed Bennamoun

PII: S0167-6393(15)30078-9
DOI: 10.1016/j.specom.2017.01.005
Reference: SPECOM 2432

To appear in: *Speech Communication*

Received date: 30 November 2015
Revised date: 1 November 2016
Accepted date: 3 January 2017

Highlights

- Develop a novel cascade feature extraction method for audio-visual speech recognition
- Firstly show the depth visual information can significantly boost visual speech recognition
- Firstly experimentally reveal different characteristics of grey and depth visual features
- Introduced the first large-scale audio-visual speech corpus that contains depth information.

A Cascade Gray-Stereo Visual Feature Extraction Method For Visual and Audio-Visual Speech Recognition

Chao Sui[a], Roberto Togneri[b], Mohammed Bennamoun[a]

[a] *School of Computer Science and Software Engineering, University of Western Australia, Perth, WA, 6009, Australia. E-mail: lukesui@hotmail.com; mohammed.bennamoun@uwa.edu.au.*

[b] *School of Electrical, Electronic and Computer Engineering, University of Western Australia, Perth, WA, 6009, Australia. E-mail: roberto.togneri@uwa.edu.au.*

Abstract

Although stereo information has been extensively used in computer vision tasks recently, the incorporation of stereo visual information in Audio-Visual Speech Recognition (AVSR) systems and whether it can boost the speech accuracy still remains a largely undeveloped area. This paper addresses three fundamental issues in this area: 1) Will the stereo features benefit visual and audio-visual speech recognition? 2) If so, how much information is embedded in stereo features? 3) How to encode both planar and stereo information in a compact feature vector? In this study, we propose a comprehensive study on the characteristics of both planar and stereo visual features, and extensively analyse why the stereo information can boost the visual speech recognition. Based on the different information embedded in planar and stereo features, we present a new Cascade Hybrid Appearance Visual Feature (CHAVF) extraction scheme which successfully combines planar and stereo visual information into a compact feature vector, and evaluate this novel feature on visual and audio-visual connected digit recognition and isolated phrase recognition. The results show that stereo information is capable of significantly boosting the speech recognition, and the performance of our proposed visual feature outperforms the other commonly used appearance-based visual features on both the visual and audio-visual speech recognition tasks. Particularly, our proposed planar-stereo visual feature yields approximately 21% relative improvement over the planar visual feature. To the best of our knowledge, this is the first paper that extensively evaluates the different characteristics of planar and stereo visual features, and we first show that using the stereo feature along with the planar feature can significantly boost the accuracy on a large-scale audio-visual data corpus.

Keywords: Audio-visual speech recognition, planar-stereo visual information, hybrid-level visual feature

1. Introduction

Speech has long been acknowledged as one of the most effective and natural means of communication between human beings. In recent decades, continuous and substantial progress has been made in the development of Automatic Speech Recognition (ASR) systems. However, in most practical applications the accuracy of ASR systems is negatively affected by exposure to noisy environments. Research on audio-visual speech recognition has been undertaken to overcome the recognition degradation that occurs in the presence of acoustic noise [1]. Despite the promising application perspective, state-of-the-art audio-visual speech recognition systems still cannot achieve adequate performance in practical applications, because most Visual Speech Recognition (VSR) systems use grey or colour information which is highly sensitive to a number of variables, such as varying illumination and head poses.

Given the limitations of the current texture information based VSR systems, exploiting stereo information can be an effective option in overcoming these challenges. Furthermore, with the availability of affordable stereo cameras, the utilisation of 3D information has led to some great successes in the computer vision community [2, 3, 4]. However, using 3D visual information on VSR to boost speech recognition accuracy has not been sufficiently studied. Motivated by the encouraging performance of 3D computer vision based methods, this paper proposes a visual and audio-visual speech recognition system that combines planar and stereo information to boost visual and audio-visual speech accuracy.

This paper makes the following two major contributions:

- First, it develops a novel cascade feature extraction method that can effectively encode both planar and stereo visual information into a compact visual feature. Moreover, this new visual feature carries both global and local appearance-based visual information, and our paper shows that both local and global appearance-based visual information is able to contribute to speech recognition. Experimental results demonstrate that the performance of this proposed feature extraction method achieves promising accuracy in different speech recognition tasks.

- To the best of our knowledge, this is also the first comprehensive work that experimentally demonstrates the efficacy of stereo visual features for speaker-independent continuous speech recognition on a large-scale (162 speakers) audio-visual corpus. The experimental results also demonstrate an improved performance with the integration of planar features. Since using stereo visual information for VSR is still a largely undeveloped area but has promising potential applications, this paper is expected to provide the community a new perspective to overcome the limitations of planar visual speech features.

The rest of this paper is organised as follows: Section 2 introduces some related works. Section 3 introduces two widely used appearance-based feature extraction methods, followed by an introduction of our proposed feature extraction scheme. The system performance is extensively evaluated in Section 4. Finally, Section 5 sets out the relevant conclusions that can be drawn from this research.

2. Related Works

In this section, a brief and up-to-date overview of some recent works relating to visual and audio-visual speech recognition is presented. A more comprehensive review can be found in [1, 5, 6].

The visual features used in visual and audio-visual speech recognition can be divided into three main categories: i) lip appearance-based features; ii) lip shape-based features; and iii) lip motion-based features [7]. In most visual and audio-visual speech recognition systems, lip motion-based features work collaboratively with the first two types of visual features to represent both temporal and spatial information of the lips.

Appearance-based visual feature extraction methods usually consider the whole lip or the lower face region as the most informative region for visual speech recognition. Among appearance-based visual features, the Discrete Cosine Transform (DCT) is the most widely used [8, 9, 10]. In terms of the utilisation of depth information, Galatas *et al.* [11, 12] have conducted encouraging pioneering research that employs depth DCT information for visual and audio-visual speech recognition. However, the integration of the depth and planar DCT features did not show significant improvement over the planar DCT features. Wang *et al.* [13, 14] also conducted a similar research using 3D data acquired using a Kinect. However, these works used a small audio-visual data corpus limited in both speaker number and speech content. Furthermore, the differences between stereo and planar features were not analysed. Given the current limitations of the stereo based visual speech recognition research, in our work, we analysed the different characteristics of planar and stereo visual features, and we propose a new feature combination method that can integrate both planar and stereo features into a compact feature vector, and show that the use of stereo information is able to significantly boost speech accuracy on a large-scale audio-visual data corpus.

In addition to DCT features, Local Binary Pattern (LBP) features have also been widely used in the computer vision community. This feature representation method has been shown to boost accuracy on various computer vision tasks [15, 16]. Based on the success of the LBP feature, Zhao *et al.* [17] introduced an LBP based spatio-temporal visual feature, called LBP-Three Orthogonal Planes (LBP-TOP), and this feature achieved impressive results in both speaker-independent and speaker-dependent visual speech recognition tasks.

Given the great success of the LBP-TOP features, numerous graph based methods have been proposed in recent years [18, 19, 20, 21, 22, 23]. These methods were able to non-linearly map the original LBP-TOP feature to a more compact and discriminative feature space and achieved very promising classification results. However, it should be noted that none of these graph-based methods have been tested for continuous speech recognition. Zhou et al. [24] reported that their method achieved promising results on classifying visemes; however, it is still unclear whether the graph based method can be used for continuous speech recognition. On the other hand, this paper proposes a method that embeds LBP-TOP features into a compact feature vector and experimentally shows that it is effective for continuous speech recognition tasks.

For lip shape-based features, the lip contours of speakers are first extracted from an image sequence and a parametric or statistical lip contour model is obtained. The parameters of the lip model are then used as visual features; typical methods used in this category include the Active Contour Model (ACM), the Active Shape Model (ASM) and the hybrid appearance and shape model, i.e., Active Appearance Model (AAM) [25, 26]. These shape-based visual features are able to explicitly capture the shape variations of the lips; however, it should be noted that the training process of the shape-based features is very time-consuming, as a large number of lip landmarks need to be laboriously labelled. This manual labelling process becomes infeasible when the corpus has a large number of recordings. Furthermore, speech accuracy degrades if the model is not appropriately and sufficiently trained [1]. Conversely, appearance-based feature extraction methods are computationally efficient and do not require any training processes. Thus, appearance-based methods are more suitable for robust visual and audio-visual speech recognition tasks with a large number of speakers. Given the advantages of the appearance-based visual features mentioned above, a new framework is presented in this paper that successfully includes both planar and stereo appearance-based features. This approach is shown to boost speech accuracy for visual and audio-visual speech recognition systems.

3. Visual Feature Extraction

In this section, two of the most widely used appearance-based visual features of recent years are reviewed [8, 9, 10, 17, 18, 19, 20, 27, 24], (i.e., DCT and LBP-TOP). Then, the different information types embedded in these two appearance-based visual features are discussed. This provides a justification for the motivation behind the proposed approach. Motivated by the different characteristics of appearance-based features, a proposed Cascade Hybrid Appearance Visual Feature (CHAVF) is also introduced in this section.

3.1. DCT

The DCT has been widely used in many visual and audio-visual speech recognition systems, as it can preserve speech relevant information in a feature vector of low dimension. In this study, this is applied to a sequence of frames of the mouth region. The DCT definition of one frame of the mouth region video is given by:

$$K(i,j) = \sum_{y=0}^{N-1} \sum_{x=0}^{N-1} f(x,y) \cos\Big(\frac{\pi(2y+1)j}{2N}\Big) \cos\Big(\frac{\pi(2x+1)i}{2N}\Big), \tag{1}$$

for $i, j, x, y = 0, 1, 2, ..., N-1$, where N is the width and height of the mouth ROI. The function $f(x,y)$ is the planar and stereo intensity values of the mouth ROI.

To reduce the computational cost and retain feature discrimination, 32 low-frequency DCT coefficients were selected in a Zig-Zag left to right scanning pattern. Each of these 32 DCT coefficients lie in the even columns of the DCT images, due to the lateral symmetry of the mouth region [28]. The 32 first and 32 second temporal derivatives were computed to capture the dynamic information of the utterances. For both planar and stereo DCT features, these static and dynamic features were used to constitute a 96 dimensional feature vector to represent the speech-related information. Furthermore, a feature mean normalisation at an utterance-level was used to compensate for illumination variations.

3.2. LBP-TOP

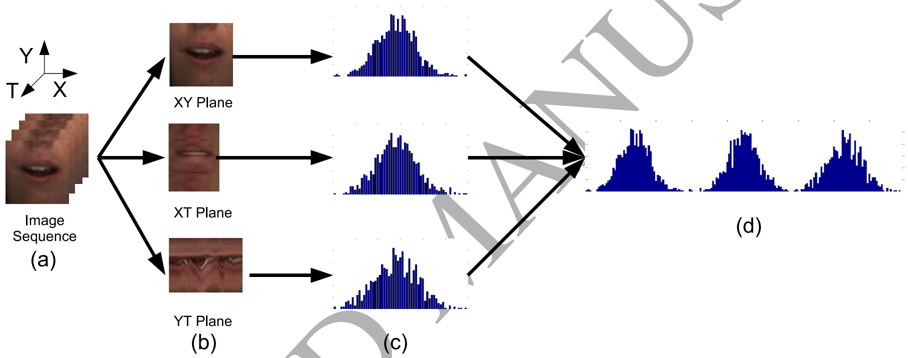


Figure 1: Lip spatio-temporal feature extraction using the LBP-TOP feature extraction. (a) Lip block volumes; (b) Lip images from three orthogonal planes; (c) LBP features from three orthogonal planes; (d) Concatenated features for one block volume with the appearance and motion.

As an alternative to DCT, Zhao *et al.* [17] introduced a spatio-temporal local texture feature extraction based on LBP and used it for visual speech recognition. This feature extraction method extracts LBP information from both the spatial and the temporal domains (i.e., Three Orthogonal Planes (TOP)). It is referred to as the LBP extracted from TOP or LBP-TOP.

Unlike the basic LBP for static images, LBP-TOP extends feature extraction to the spatio-temporal domain, which makes LBP an effective dynamic texture descriptor. Given the 2D spatial coordinates and the temporal coordinates, X, Y at time, T, a histogram is generated to accumulate the presence of different binary patterns across the XY , XT and YT planes (see Figure 1). Once the LBP histograms are generated from the three planes, a feature vector is constructed by concatenating the three histograms to represent both the lip appearance and its motion. Uniform patterns [29] were used in this study to reduce the dimension of the LBP-TOP feature vector, and the dimensionality of the LBP-TOP feature is 177.

To improve speech recognition performance, the mouth region was further divided into several subregions, as elaborated in [17], to extract LBP-TOP features from each subregion and concatenate the respective features. We experimentally found that the mouth image needs to be divided into 2×5 regions (see Figure 2) to achieve the best results. Hence, extracting the planar and stereo LBP-TOP features from the 2×5 regions results in a 1770-dimensional feature ($177 \times 2 \times 5$), and a 3540-dimensional hybrid feature is formed after concatenating planar and stereo features.

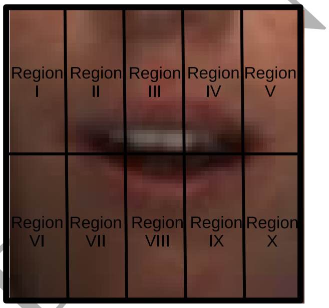


Figure 2: The mouth region is divided into 10 subregions.

3.3. Cascade Hybrid Appearance Visual Feature Extraction

Comparing two most commonly used appearance features (i.e., DCT and LBP-TOP) introduced above, one can note that these two feature methods extract features from two different information representation perspectives. As detailed in Eq. 1, each component $Y(i, j)$ of the DCT feature is a representation of the entire mouth region at a particular frequency. Thus, the DCT is a global feature representation method. Conversely, the LBP-TOP feature uses a descriptor to represent the local information in a small neighbourhood. Therefore, the LBP-TOP is a local feature representation. Also, as experimentally analysed in Section 4, these two types of features carry different kinds of information. Therefore, finding a way to embed both global and local information into a compact visual feature vector should achieve better speech accuracy compared to the individual application of these two widely used visual features. However, simply concatenating these two types of features is not practical for speech recognition, as the feature vector would be too large and make the system succumb to the curse of dimensionality. Hence, a feature dimension reduction process is essential to represent both local and global information in a single compact feature vector. The above analysis motivated the development of a cascade feature extraction framework that was able to combine both global and local information using a compact feature vector.

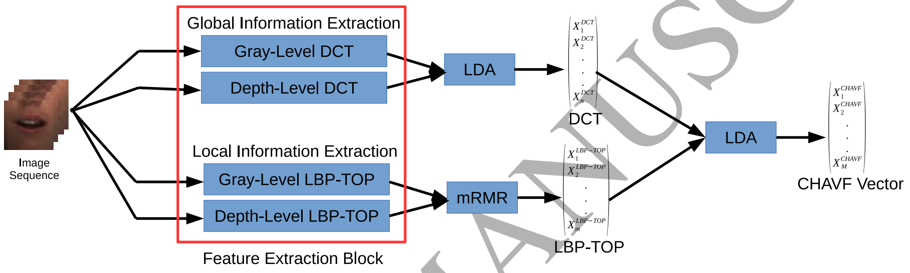


Figure 3: Cascade Hybrid Appearance Visual Feature (CHAVF) extraction.

Figure 3 provides an overview of the proposed system. In the first stage, the mouth image sequence is fed into the feature extraction block that consisted of two parallel procedures; that is, global appearance-based and local appearance-based feature extraction. The global appearance-based feature extraction uses DCT to preserve speech relevant information with a 192 dimensional feature vector (see Section 3.1). LDA is then used to reduce the dimensionality of the DCT feature vector.

However, it was found that LDA usually fails to obtain a proper transformation for feature reduction when the raw 3540-dimensional LBP-TOP feature was applied on large-scale continuous speech recognition. Consider the continuous digit sequence recognition introduced in Section 4; in this experiment, HMMs were employed to model each of the 11 digits (two pronunciations for zero). Thus, the total target classification for LDA was 330 (i.e., 30 × 11). However, modelling LDA on 3540-dimensional features for 330 classes would require an extremely large amount of video data. The data corpus used in this work is one of the largest digit sequence corpus available; however, the amount of data (approximately 300,000 frames) was still insufficient for LDA modelling.

Besides the conventional LDA, a kernel LDA [30] that nonlinearly maps the original LBP-TOP features to a feature space using the kernel trick could solve this insufficient data problem, a kernel LDA is computationally expensive and, in this case, would have been intractable, as the number of training examples would have been too large for a kernel LDA to process. Yu and Yang [31] proposed a direct LDA to overcome the training difficulty when the number of training samples is smaller than the feature dimension. In our work, we compared our proposed method with the direct LDA based method, and report the results in Section 4.2.

In addition to LDA and its variants, in recent years, numerous graph-based feature reduction methods for LBP-TOP have been proposed [18, 19, 20, 21, 22, 23]; however, all of these works focus on the speech classification problem that represents a simpler data reduction problem than the one used for the continuous speech recognition task. Furthermore, the high computational requirements limit its applicability to this task. Fortunately, Gurban et al. [8] proposed less computationally demanding methods, called Mutual Information Feature Selectors (MIFS) for visual speech recognition. The MIFS for LBP-TOP feature reduction was chosen for two reasons. First, MIFS has a strong theoretical justification that comes from the Fanos inequality [32]. The Fano's inequality gives a lower bound probability of error in the system. Thus, a feature with high mutual information of classes is more helpful for classification. Second, the MIFS is computationally efficient because no complex training process is required. Thus, given the amount of data in the corpus and the task being performed, MIFS is an effective option for solving this problem, as it creates a good balance between the quality of features and the computational time.

Among the different types of MIFSs, the simplest method for selecting feature components is the Maximum Mutual Information (MMI). Let $\mathbf{x}_M^i$ be the visual feature of the frame i $(i = 1, 2, ..., n)$ in the M-dimensional space, S_M (in our system, M is the dimension of the raw LBP-TOP feature, i.e., 3540). Here n is the total number of frames over the entire collection of video sequences used for training. A feature subset S_k ($k \leq M$, in this work, $k = 310$) is selected using MMI:

$$S_k = S_{k-1} \cup \{ \underset{\mathbf{x}_j \in (S_M \setminus S_{k-1})}{\arg\max} \; I(\mathbf{x}_j; C)\}, \tag{2}$$

where C is one of the 330 classes used in this work, and $\mathbf{x}_j \in (S_M \setminus S_{k-1})$

means that $\mathbf{x}_j$ is in the feature space S_M, but does not belong to the subset S_{k-1}. The mutual information $I(\mathbf{x}_j; C)$ can be estimated as:

$$I(\mathbf{x}_j; C) = \sum_{x \in \mathbf{x}_j} \sum_{c \in C} p(x, c) \log \frac{p(x, c)}{p(x)p(c)}, \tag{3}$$

where $p(x, c) = p(\mathbf{x}_j = x, C = c)$ is the joint probability density function of $\mathbf{x}_j$ and C. In our system, a histogram with 100 bins is used to compute the mutual information. Therefore, $p(x)$ can be estimated using the total number of training samples and the total number of training samples that falls into the interval $\lfloor x \rfloor \leq x \leq \lceil x \rceil$.

Obviously, MMI (Eq. 2) finds the k most informative feature dimensions from the original space in the information theoretic sense. However, it does not necessarily follow that this set of k feature components is the most informative feature set for target classification; rather, this feature set could have a rich redundancy. Peng et al. [33] proposed another type of MIFS called minimal-Redundancy-Maximal-Relevance (mRMR) that takes the feature redundancy into consideration:

$$S_k = S_{k-1} \cup \{ \underset{\mathbf{x}_j \in (S_M \setminus S_{k-1})}{\arg\max} [I(\mathbf{x}_j; C) - \frac{1}{k-1} \sum_{\mathbf{x}_l \in S_{k-1}} I(\mathbf{x}_j; \mathbf{x}_l)]\}, \tag{4}$$

where $I(\mathbf{x}_j; \mathbf{x}_l)$ is the mutual information between the feature components $\mathbf{x}_j$ and $\mathbf{x}_l$, which can be calculated using Eq. 5. $I(\mathbf{x}_j; \mathbf{x}_l)$ acts as a penalty term to approximate the feature redundancy between different feature components.

$$I(\mathbf{x}_j; \mathbf{x}_l) = \sum_{x_1 \in \mathbf{x}_j} \sum_{x_2 \in \mathbf{x}_l} p(x_1, x_2) \log \frac{p(x_1, x_2)}{p(x_1)p(x_2)}, \tag{5}$$

Another MIFS which has been widely used is the Conditional Mutual Information (CMI) [34]. It uses the relevant redundancy between features when the class labels are given:

$$S_k = S_{k-1} \cup \{ \underset{\mathbf{x}_j \in (S_M \setminus S_{k-1})}{\arg\max} [I(\mathbf{x}_j; C) - \max_{\mathbf{x}_l \in S_{k-1}} I(\mathbf{x}_j; \mathbf{x}_l | C)]\}, \tag{6}$$

where the penalty term $I(\mathbf{x}_j; \mathbf{x}_l | C)$ takes the class into account, and is given by:

$$I(\mathbf{x}_j; \mathbf{x}_l | C) = \sum_{x_1 \in \mathbf{x}_j} \sum_{x_2 \in \mathbf{x}_l} \sum_{c \in C} p(x_1, x_2, c) \log \frac{p(c)p(x_1, x_2, c)}{p(x_1, c)p(x_2, c)}, \tag{7}$$

Comparing Eq. 4 with Eq. 6, the CMI appears to be more pertinent to the classification task than mRMR, as the penalty term of CMI would be relevant to the classification task. However, calculating the penalty term $I(\mathbf{x}_j; \mathbf{x}_l; C)$ would require a much larger amount of data compared to the estimation of the penalty term $I(\mathbf{x}_j; \mathbf{x}_l)$ of mRMR [8].

As explained in Section 4, it was found that mRMR performs much better on the LBP-TOP feature than MMI and CMI. Consequently, mRMR was chosen as the feature selector for the proposed CHAVF. After the extraction of the relatively compact DCT and LBP-TOP features, their concatenation remained very large for the classifier to process. Thus, LDA was again used to further reduce the dimensionality of the features. With the application of the proposed novel cascade feature extraction framework, information representing different characteristics (i.e., global and local) and modalities (i.e., texture and stereo) was successfully embedded into a compact feature vector. To the best of our knowledge, this is the first compact visual feature type that can represent speech relevant information from multiple information representation aspects.

4. Performance Evaluation and Results

Two data corpora were used in this work. The AusTalk (see Section 4.1.1) was used for speaker-independent visual speech recognition (Section 4.2) and speaker-independent audio-visual speech recognition (Section 4.4). The OuluVS (see Section 4.1.2) was used for the speaker-independent visual phrase classification experiments (Section 4.3)

4.1. The AusTalk and OuluVS Corpora

4.1.1. AusTalk

Since this paper mainly focuses on how to employ stereo visual information for speech recognition, a suitable corpus that contains stereo data needs to be used. In terms of existing corpora which contain stereo data, the data from AV@CAR [35] only contains stereo facial expression information and does not have any 3D visual speech related content. Hence, it cannot be used for stereo visual speech recognition. Although the WAPUSK20 [36] and the AVOZES [37] contain 3D speech data, these data corpora are limited in either the number of speakers or speech content. Hence, we used a recently developed data corpus that addresses these limitations.

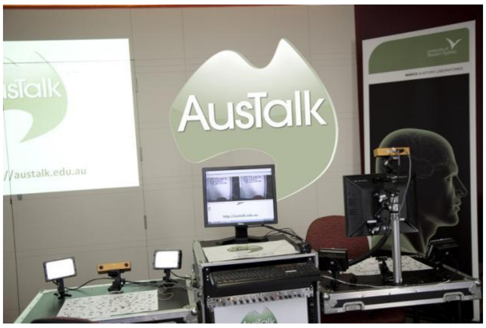


Figure 4: The recording environments and devices used to collect the AusTalk data.

The main data corpus used in this paper was collected by an Australia wide research project called AusTalk, funded by the Australian Research Council [38, 39, 40, 41]. This research project involved more than 30 scientists from 11 Australian universities and resulted in the creation of a large-scale data corpus that can be used in audio-visual speech recognition research. The AusTalk corpus consists of a large 3D audio-visual database of spoken Australian English recorded at 15 different locations in each of Australias states and territories; the contemporary voices of 1,000 Australian English speakers of all ages were recorded to capture variations in accents, linguistic characteristics and speech patterns. To satisfy a variety of speech driven tasks, several types of data was recorded, including isolated words, digit sequences and sentences.

To collect the AusTalk data, a Standard Speech Science Infrastructure Black Box (SSSIBB) was designed [40]. The recording equipment includes head-worn and desktop microphones, digital audio acquisition devices and a stereo camera (see Figure 4). A Bumblebee stereo camera, mounted approximately 50cm from the speakers, was used to collect the 3D information of the speakers in addition to the texture (RGB) information. Although the accuracy of this stereo camera is not as high as some more expensive cameras used in other 3D driven tasks (e.g., those used by the 3dMD company [42] for precise surface imaging based applications), it is a low cost stereo camera which can be used in a much wider range of real-life applications. The details for the extraction of the audio and visual features can be found in [41] and the configuration parameters of the Bumblebee camera are listed in Table 1. Figure 5 represents a few video data samples.

To generate the required planar-level and stereo-level information, face

Table 1: Bumblebee camera configuration used for building our system.

Attribute	Value
Resolution	640 × 480
Disparity Range	$[41, 137]$
Stereo Mask	11
Edge Correlation	On
Edge Mask	7
Sub-pixel Interpolation	On
Surface Validation	On
Surface Validation Size	400
Surface Validation Difference	On
Uniqueness Validation	On
Uniqueness Validation Threshold	1.44
Texture Validation	On
Texture Validation Threshold	0.4
Back Forth Validation	On

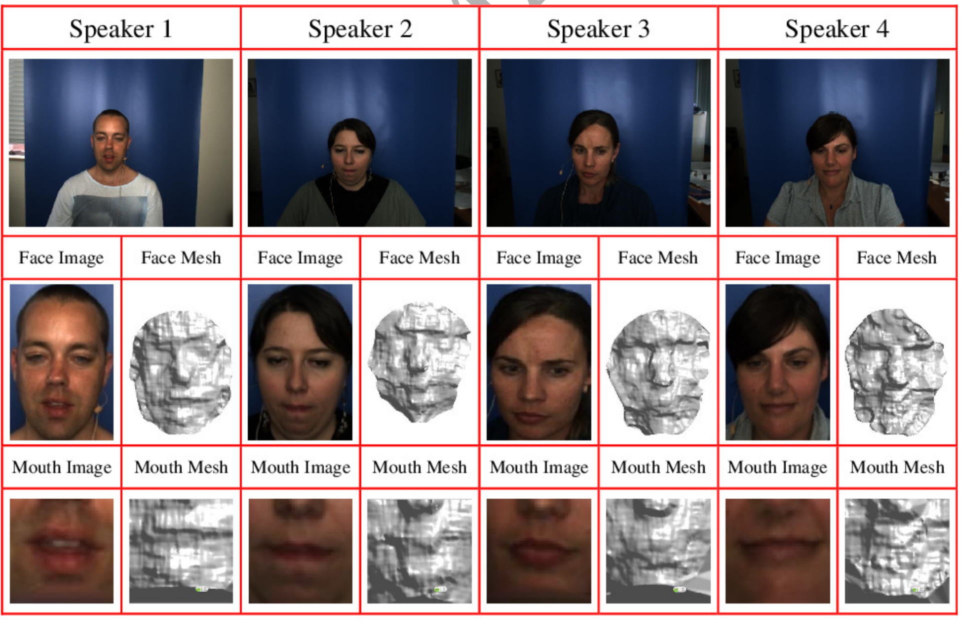


Figure 5: Sample RGB images and meshes from the AusTalk visual dataset.

detection is first performed, and the face ROIs are cropped from the original stream using the Haar features and Adaboost [43]. In order to align the face to the same position and to correct the head pose, the Iterative Closest Point (ICP) algorithm [44] is applied on the depth data. To reduce the computational load of the face alignment, only the upper faces are used, because the upper face can be considered as rigid and is therefore less affected by facial expressions [45]. The upper face of the first frame is used as the reference model, and the upper faces of the remaining frames are registered to the reference model. Then, a cubic interpolation is performed to fill in holes and reduce noise (e.g., spikes). Given the aligned point cloud faces, the mouth region can then be easily cropped by applying the Haar features and Adaboost. We use a square of $50mm \times 50mm$ centred at the mouth centre to crop the mouth. Since there is a one-to-one correspondence between the points in the face point cloud and the pixels in its corresponding texture image, the 2D mouth images (50×50) can be extracted in this step as well.

To show the effectiveness of the proposed method in a large-scale, speaker independent, continuous speech recognition task, experiments were conducted using the digit sequence session of the AusTalk data corpus. In this session, 12 4-digit sequences were collected from each of the speakers (see Table 2). For digit '0, two pronunciations (i.e., zero and oh) were used to capture the different speech habits. This set of digit strings was carefully designed to ensure that each digit (i.e., 0-9) occurred at least once in each serial position. Take digit '1' for example, as listed in Table 2, it occurs in the 4th recording '123z' (in the first position), the 1st recording 'z123' (in the second position), the 10th recording '3z12' (in the third position), and the 7th recording '23z1' (in the fourth position). The digits in the data corpus were read in a random manner without any unnatural pause to simulate PIN recognition and telephone dialing tasks. This configuration differs to the popular audio-visual data corpus like CUAVE [46] that reads digits in a sequential manner with a relatively long pause between digits. The random digit sequences that were recorded in AusTalk made the recognition task more difficult; moreover, this configuration also ensured that the digit recording of the data corpus was more balanced. To capture any within-speaker variability over time, each participant speaker was encouraged to attend two separate recording sessions. There was at least a one-month gap between the two recording sessions; however, the speech contents recorded in these two sessions were identical. As not all speakers attended a second session, some recordings only contain data from the first session. In this study, data recordings from

162 speakers (around 1,900 utterances) were used. Of these 162 speakers, 148 speakers attended one recording session and 14 speakers attended both recording sessions. However, a small number of recordings were not used for speech recognition due to several technical issues. For example, the speakers involuntarily moved their head and body during the recording, so that the stereo camera failed to extract useful depth information. Given the large amount of data that we used in this research, it is not feasible to manually crop the mouth region from the image. Hence, an Adaboost based face and mouth detection algorithm was used, and in some cases the face and mouth detection failed to detect the mouth from the videos. Like some other works that have had similar issues [17, 18, 19, 20, 24], these recordings are removed from the data corpus, and 1861 (out of 1992) recordings were used in our experiments.

Table 2: Digit sequences in the AusTalk data corpus. For the digit '0', there are two possible pronunciations: 'zero' ('z') and 'oh' ('o').

No.	Digits	No.	Digits	No.	Digits
01	z123	02	942o	03	6785
04	123z	05	7856	06	2o94
07	23z1	08	49o2	09	8567
10	3z12	11	5678	12	0429

In the proposed work, the Hidden Markov Toolkit (HTK) [47] was used to implement HMMs for digit sequence recognition. In this experiment, the digit recognition task was treated as a connected word speech recognition problem with a simple syntax (i.e., any combination of digits and silence was allowed in any order). With respect to the HMM model, 11 word models were used with 30 states to model 11 digit pronunciations, a 5-state HMM was used to model the silence of the beginning and the end of the recording, and a 3-state HMM was used to model the short pause between the digit utterances. Each HMM state was modelled by nine Gaussian Mixtures with diagonal covariance. In relation to the experimental setup, the 162-speaker digit data recordings were divided into 10 groups; the speakers in the different groups did not overlap. A 10-fold cross validation was then employed to increase the statistical significance of the results. The average speech accuracy of 10 runs was then reported.

4.1.2. OuluVS

In recent years, the majority of high-quality work in this area has focused on speech classification [6]. OuluVS [17] was used to compare the proposed CHAVF with other state-of-the-art systems. OuluVS, a widely used data corpus that performs speech classification, is a visual-only data corpus comprising of 10 English phrases (see Table 3) uttered by 17 male speakers and 3 female speakers. The data in this corpus was collected using a SONY DSR-200AP 3CCD-camera with a frame rate of 25 fps. Each phrase was repeated nine times by each speaker. As in Zhao et al. work [17], 817 sequences from 20 speakers were used in our experiments. The second degree polynomial kernel Support Vector Machine (SVM) was used as the classifier. This is the same classifier as the one used in [17, 18]. In terms of the verification process, a leave-one-speaker-out approach was adopted, such that the recordings of 19 speakers was used for the training dataset and the left out speaker was used to test the data for each of the 20 runs.

Table 3: The 10 phrases in OuluVS data corpus.

No.	Phrase	No.	Phrase
01	Excuse me	02	Goodbye
03	Hello	04	How are you
05	Nice to meet you	06	See you
07	I am sorry	08	Thank you
09	Have a good time	10	You are welcome

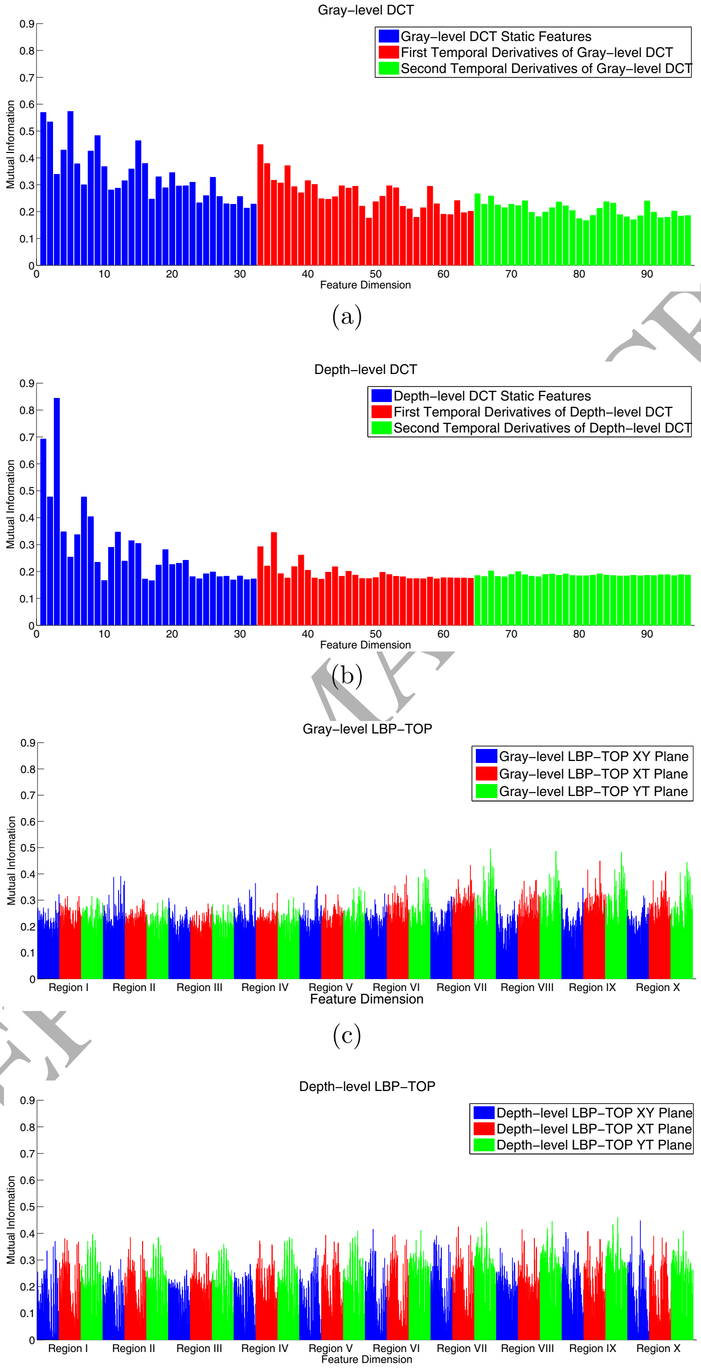
4.2. Speaker-Independent Visual Speech Recognition

To examine the amount of relevant information for planar and stereo visual features, the mutual information (see Eq. 3) of the different visual components are plotted in Figure 6. In relation to the DCT features, Figures 6a and 6b show that the amount of information carried by planar and stereo is quite different. This is an interesting observation firstly revealed by this work, and explains why the integration of planar and stereo features by our proposed method is capable of boosting speech recognition accuracy. Furthermore, these two kinds of visual features also share a common characteristic. The DCT static coefficients had more discriminative information about the classes (i.e., the states of digit HMM models) than the dynamic coefficients. Previous studies found similar results for the hVd words recognition from a linear discriminative perspective [48]. The hVd words are a set of words starting with the letter 'h', a vowel in the middle and ending with the letter 'd' and they are important for acoustic-phonetic analysis. Both works show that the DCT static features were more discriminative than the dynamic ones.

In this study, it was found that the mutual information of feature components that contribute to both visual and audio-visual speech recognition is usually larger than 0.3. Figure 6a and Figure 6b show that the number of stereo DCT components that have considerably large mutual information I ($I \geq 0.3$) was much smaller than the planar counterpart (31 vs. 10). The difference between the planar and the stereo features does not necessarily mean that the planar features were more informative than the stereo features, as high information redundancy may exist in the planar features.

For the LBP-TOP features, the mutual information of each mouth subregion (see Figure 2) has been listed in Figures 6c and 6d. From these figures, it can be observed that the planar LBP-TOP components of the lower mouth regions (i.e., region VI to region X in Figure 2) were more informative than the components of the upper mouth regions (i.e., region I to region V in Figure 2). This is consistent with the observation that in human speech production most movements for talking occur in the lower lip and the jaw. An experiment on LBP-TOP features was also carried out (see Table 4 for the results). As displayed in Table 4, the accuracy was about 10% higher after the mouth region was divided into 2×5 blocks. Thus, the mouth region subdivision scheme introduced in Section 3.2 boosted speech accuracy. Interestingly, unlike the DCT feature introduced above, for the LBP-TOP features, the temporal features (i.e., the features extracted from the XT and YT planes) were generally more informative than the spatial features (i.e., the features extracted from the XY plane). These complementary behaviours of DCT and LBP-TOP features explain why our proposed method is effective, as our method is able to automatically obtain more static speech-relevant information from the DCT, while incorporating more dynamic information from the LBP-TOP.

Figure 7 shows the visual-only speech recognition results using the hybrid-level, planar (grey-level) and stereo (depth-level) DCT and LBP-TOP features with LDA and the MIFS selection methods (i.e., MMI, mRMR and CMI). For the DCT visual features, the use of LDA achieved the highest accuracy for planar, stereo and hybrid-level features (i.e., 54.66%, 55.19%, and 64.93%, respectively). Also, it is interesting to note that with the application of LDA, the accuracy achieved by the stereo DCT feature was almost



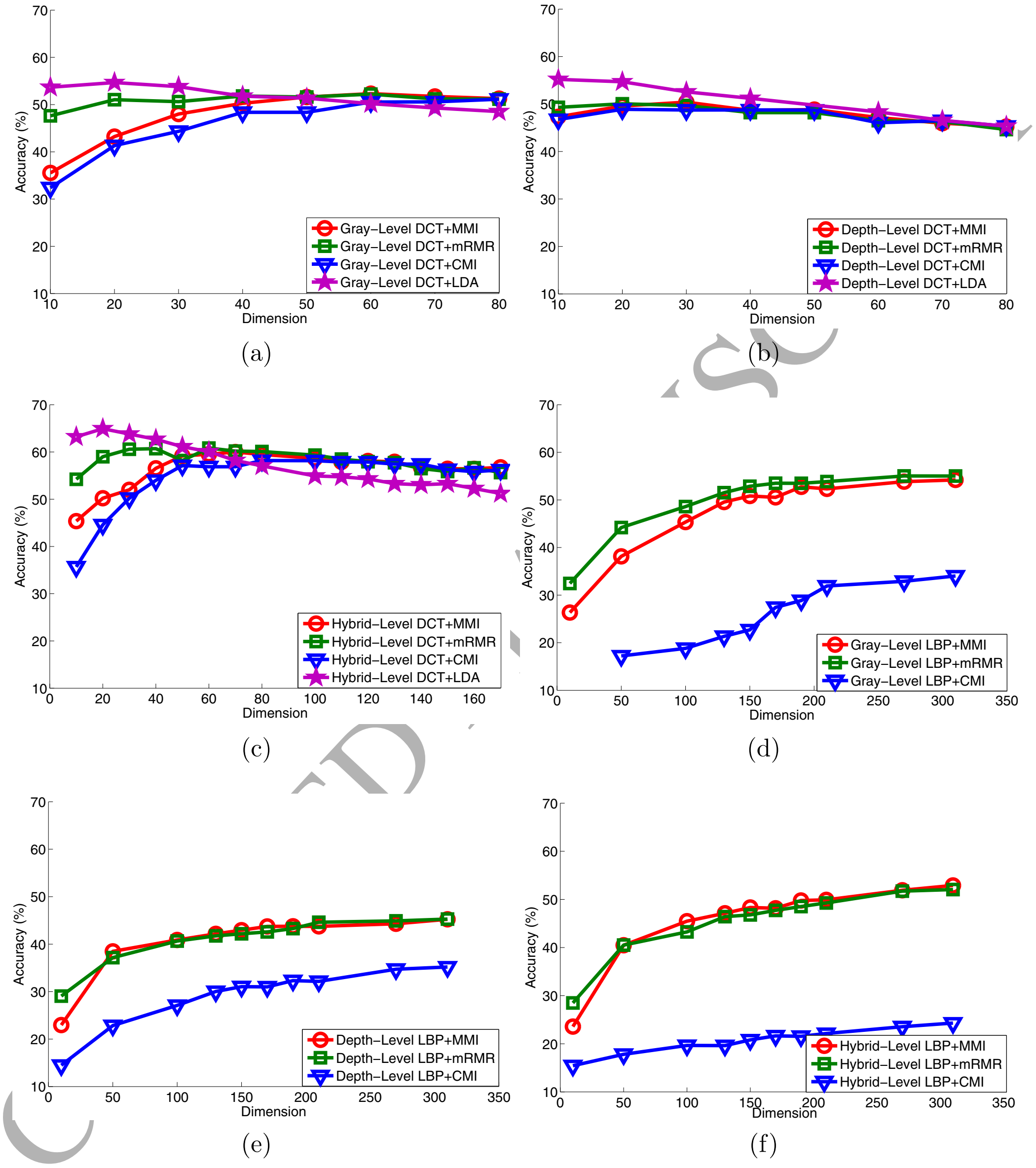
(a)

(b)

(c)

(d)

Figure 6: The comparison of the amount of relevant information ($I(\mathbf{x}_j; C)$) embedded in different types of feature dimensions.



(a) (b)

(c) (d)

(e) (f)

Figure 7: The performance of visual-only speech recognition using various feature types and feature reduction techniques: planar (gray-level) DCT (Figure 7a), stereo (depth-level) DCT (Figure 7b), hybrid-level DCT (Figure 7c), planar (gray-level) LBP-TOP (Figure 7d), stereo (depth-level) LBP-TOP (Figure 7e), hybrid-level LBP-TOP (Figure 7c).

Table 4: Planar LBP-TOP feature. The superscript 1×1 which represents the entire lip region are fed into the feature extraction procedure without subdivision, while the superscript 2×5 which represents the mouth region is divided into 10 subregions as shown in Figure 2. The subscript represents the radii of the spatial and temporal axes (i.e., X, Y and T) and the number of neighbouring points in these three orthogonal planes are 8 and 3, respectively.

Feature Type	Feature Extraction	Visual Speech Accuracy
Planar	$LBP - TOP_{8,3}^{1\times1}$	46.78%
	$LBP - TOP_{8,3}^{2\times5}$	53.06%
Stereo	$LBP - TOP_{8\_3}^{1\times1}$	35.31%
	$LBP - TOP_{8,3}^{2\times5}$	45.28%

the same as that achieved by the planar DCT feature (i.e., 55.19% versus 54.66%). This indicates that while the stereo images were quite noisy and the stereo lip regions were barely visible to human eyes (see Figure 5), the stereo DCT feature was still capable of representing relevant geometrical information. One reason for the promising accuracy yielded by the low quality stereo image sequences is that the stereo DCT is a global feature representation method that is insensitive to image noise. In relation to the hybrid DCT feature that combines both planar and stereo global information, the visual speech accuracy was approximately 10% higher than either of the corresponding planar and stereo features.

In relation to the LBP-TOP visual feature, with the application of mRMR, the 290-dimensional planar, 310-dimensional stereo and 310-dimensional hybrid-level features yielded 53.06%, 45.28% and 52.04%, respectively accuracy (see Figure 7). Note that the stereo LBP-TOP feature did not perform as well as the planar LBP-TOP feature. As illustrated in Figure 5, the quality of the stereo images was much worse than the quality of the RGB images. As the LBP-TOP is a local information representation method, it is sensitive to the noise of the image. Thus, the performance of the planar LBP-TOP feature was better than that of its stereo counterpart. As can be seen in Figure 7, the different speech recognition performance demonstrated by planar and stereo features explains the superiority of our proposed method, i.e., to boost speech recognition accuracy, our method can automatically encode more planar information using a global extraction method, while encoding less stereo information from a local perspective.

As explained in Section 3.3, after MIFS was applied to select the most informative components from the raw LBP-TOP feature, LDA was used to fur-

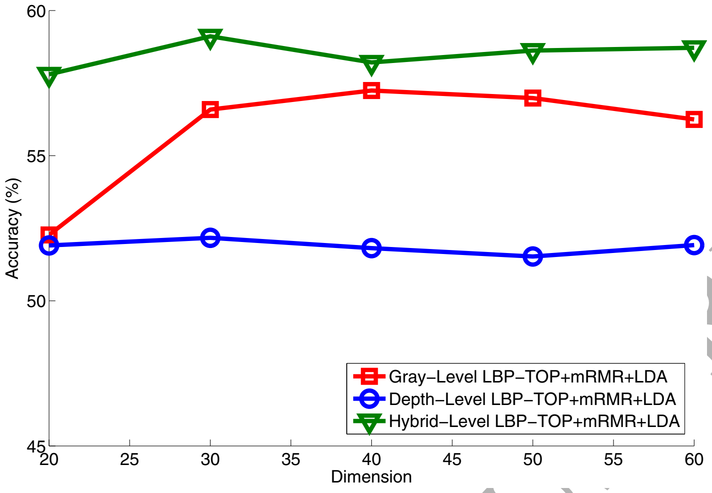


Figure 8: The visual-only speech recognition performance of the mRMR selected LBP-TOP features followed by LDA for further feature dimension reduction.

ther reduce the feature dimensionality. In relation to MIFS, mRMR achieved the highest visual speech accuracy for the hybrid-level LBP-TOP with a feature dimension of 310 as evident in Figure 7. Thus, a 310-dimensional hybrid LBP-TOP feature vector was used that was selected by mRMR for further dimensionality reduction. Figure 8 shows the visual-only speech recognition performance of the mRMR selected LBP-TOP features followed by LDA for the further feature dimension reduction. As shown in Table 5, the accuracy of the planar, stereo and hybrid-level LBP-TOP features with mRMR were 53.06%, 45.28% and 53.54%, respectively. Using LDA for further feature reduction, the visual speech accuracy for the planar, stereo and hybrid-level features were 57.23%, 52.17%, 59.11%, respectively. These results show that the novel cascade feature reduction framework proposed in this paper is very effective for visual speech recognition.

As summarised in Table 5, for the planar features, both the DCT and the LBP-TOP feature achieved promising accuracy and the LBP-TOP feature (at 57.24%) performed better than its DCT counterpart (at 54.66%). Zhao *et al.* [17] also found that the planar LBP-TOP feature was superior to the DCT feature. In relation to the stereo visual features, unlike their planar counterparts, the DCT feature outperformed the LBP-TOP feature (55.19% versus 52.17%), as the global information representation (i.e., DCT) was

Table 5: Visual speech recognition performance comparison between popular appearance features and our proposed method.

Modality	Feature	Dimension	Accuracy
Planar	DCT + MMI [8]	60	52.32%
	DCT + mRMR [8]	60	52.21%
	DCT + CMI [8]	80	51.14%
	DCT + LDA	20	54.66%
	LBP-TOP [17] + MMI [8]	290	52.46%
	LBP-TOP [17] + mRMR [8]	310	53.06%
	LBP-TOP [17] + CMI [8]	310	37.06%
	LBP-TOP + mRMR + LDA	**40**	**57.24%**
Stereo	DCT + MMI [8]	30	50.98%
	DCT + mRMR [17]	20	50.04%
	DCT + CMI [8]	20	48.89%
	DCT + LDA	10	55.20%
	LBP-TOP [17] + MMI [8]	310	44.60%
	LBP-TOP [17] + mRMR [8]	310	45.28%
	LBP-TOP [17] + CMI [8]	310	34.74%
	LBP-TOP + mRMR + LDA	**30**	**52.17%**
Hybrid	DCT + MMI [8]	70	59.98%
	DCT + mRMR [8]	50	60.82%
	DCT + CMI [8]	80	58.18%
	DCT + LDA	20	64.93%
	LBP-TOP [17] + CCA [49]	294	41.53%
	LBP-TOP [17] + MMI [8]	310	53.54%
	LBP-TOP [17] + mRMR [8]	310	54.28%
	LBP-TOP [17] + CMI [8]	310	37.06%
	LBP-TOP [17] + direct LDA[31]	40	52.32%
	LBP-TOP + mRMR + LDA	**30**	**59.11%**
	CHAVF using CCA	40	64.34%
	CHAVF	**50**	**69.18%**

more suitable to extracting information from noisy stereo images.

Obviously, the planar, the stereo DCT and LBP-TOP visual features contained considerable speech related information. Thus, combining these information sources to form a more discriminate visual feature should have boosted speech accuracy. It is clear that for both DCT and LBP-TOP features the integration of the planar and the stereo information yielded better speech accuracy as compared to any of the single modalities (see Table 5). Specifically, the accuracy obtained by the hybrid DCT feature was 64.93% (i.e., 10 % higher than the standard planar DCT feature that was 54.66%). The stereo LBP-TOP feature did not perform as well as its planar counterpart; however, the integration of the stereo local information with the planar local information still yielded a better accuracy (i.e., 59.11%) and was approximately 2% higher than the standard planar LBP-TOP feature (i.e., 57.24%). This confirms one significant aspect of this study; that is, even low quality stereo visual data can be used for speech recognition. Furthermore, visual speech accuracy significantly increases by integrating both stereo and planar visual features.

As introduced in Section 3.3, the conventional LDA cannot be used for the LBP-TOP feature because of the high dimensionality of LBP-TOP. Hence, in our experiments we compared our method with an LDA variant, i.e., the direct LDA. Our proposed cascade feature reduction scheme showed superiority over the direct LDA method (59.11% vs 52.32%). Since the direct LDA is a special case of LDA, it only works well in applications with well-separated classes [50], and therefore did not outperform our proposed method.

In relation to the proposed hybrid visual feature extraction scheme, it not only combined two separate information resources (i.e., the planar modality and the stereo modality), it also took into account two complementary information representation methods (i.e., local and global information). Thus, the proposed feature extraction framework was able to produce an even higher level of speech accuracy. The visual speech accuracy of the proposed feature (i.e., CHAVF) was 69.18% and outperformed all the listed visual features (see Table 5,). Compared with the DCT+LDA, the proposed CHAVF yielded an overall improvement of approximately 4.25% .

The major contribution of our work is a feature extraction scheme that can represent visual speech information from different views. We also compared our results with Canonical Correlation Analysis (CCA), which is one of the most popular techniques used for multi-view feature learning [51]. The same with the LDA training, introduced in Section 3.3, 330 HMM states for

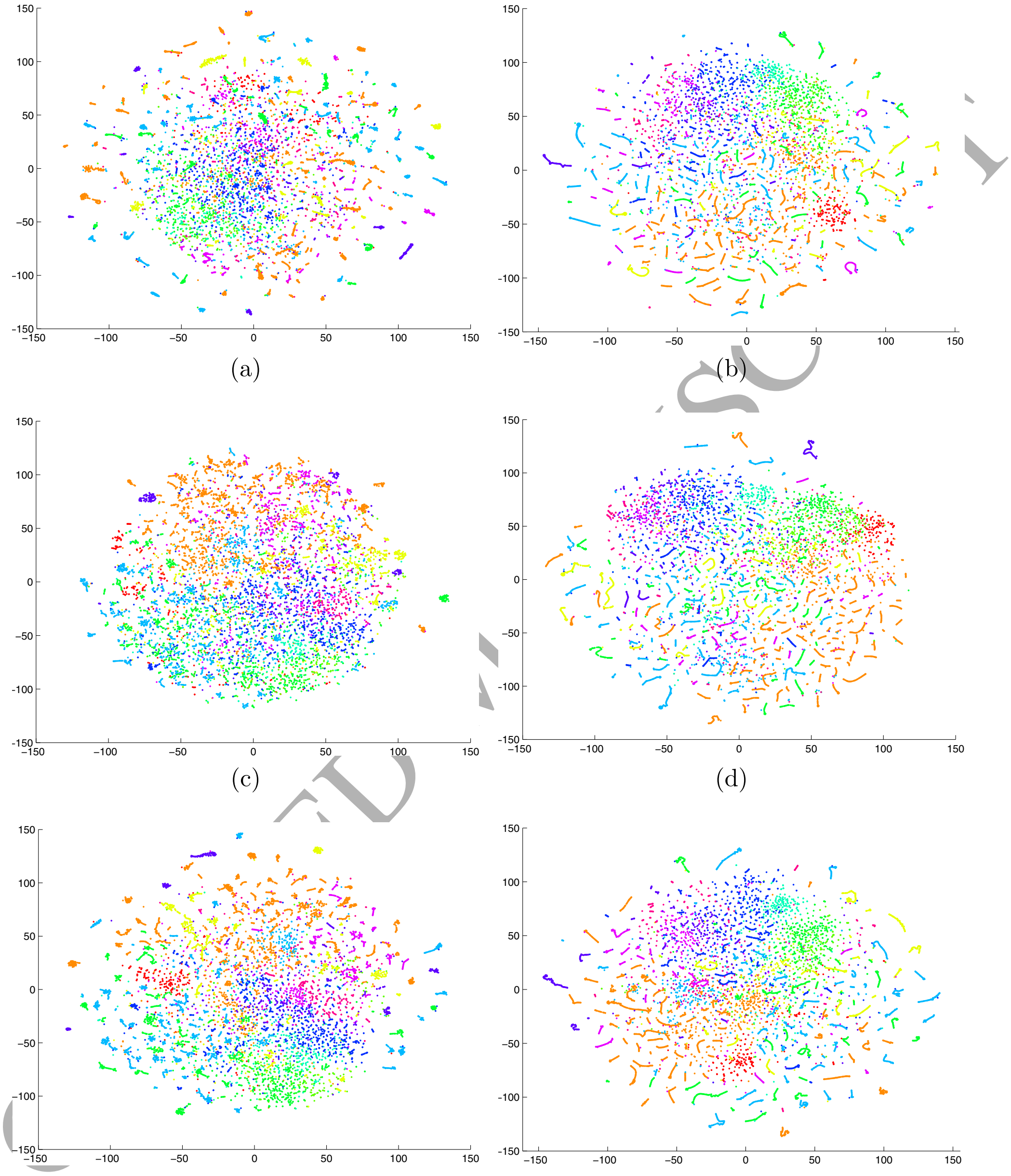
11 digits were used as targets in the classification for CCA. After employing CCA, two 20-dimensional feature vectors were produced for each of the DCT and LBP-TOP features. Concatenating these two feature vectors results in a 40-dimensional feature, which constitutes the hybrid CCA based feature. The first experiment we performed used CCA on planar and stereo LBP-TOP features to learn a hybrid LBP-TOP feature. Experimental results showed that the hybrid LBP-TOP feature with our proposed cascade feature dimension reduction scheme yielded a significantly better result compared with the CCA learning scheme (59.11% vs 41.53%). The second experiment used CCA to learn a combined local-global hybrid feature. We replaced the LDA at the second stage of our proposed scheme (as shown in Fig. 3) with CCA, and our proposed method with LDA outperformed the CCA variants (69.18% vs 64.34%).

To gain insight into why the proposed feature outperformed other visual features, a recently introduced visualisation method called t-SNE [52] was used to produce 2D embeddings of visual features. Data points close in the high dimensional feature space are also close in the 2D space produced by t-SNE. Figure 9 shows the 2D mapping of the proposed method and several features that achieved the best visual speech accuracy in their corresponding categories. The data points in Figure 9 represent video frames and different colours correspond to different classes (i.e., the different states of the HMM models). For clarity, the fifth state of each digit HMM model for visualisation was randomly chosen. Figure 9 shows the integration of planar and stereo features were more visually distinctive compared to conventional planar visual features (see Figures 9e and 9f) that exhibited more dispersion (see Figures 9a to 9d). Thus, from the data visualisation perspective, Figure 9 highlights a key finding of this study; that is, the quality of a visual feature can be improved by the integration of both stereo and planar visual features.

In addition to comparing our proposed method with other feature-level fusion methods, we also compared our method with classifier-level fusion methods. In this work, the Multi-Stream HMM (MSHMM) was used to fuse the classification results from the DCT and LBP-TOP features. The HMM training for DCT and LBP-TOP features was conducted separately. In the test step, the emission likelihood was computed as:

$$\log b_j(o_t^{fuse}) = \lambda_{dct} \log b_j(o_t^{dct}) + \lambda_{lbp} \log b_j(o_t^{lbp}), \tag{8}$$

where $b_j(o_t^{fuse})$, $b_j(o_t^{dct})$ and $b_j(o_t^{lbp})$ are the joint emission probability, the DCT stream emission probability and the LBP-TOP stream emission prob-



(a)

(b)

(c)

(d)

(e)

(f)

Figure 9: 2D t-SNE visualisation of different visual features with various feature reduction techniques. Figure 9a: Planar DCT+LDA; Figure 9b: Planar LBP+mRMR+LDA; Figure 9c: Stereo DCT+LDA; Figure 9d: Stereo LBP+mRMR+LDA; Figure 9e: Hybrid-level DCT+LDA; Figure 9f: Our proposed CHAVF.

ability, respectively. The λ_{dct} and λ_{lbp} are the weights of the DCT and the LBP-TOP streams, and $\lambda_{dct} + \lambda_{lbp} = 1$. The transition probability of the MSHMM was estimated by the weighted sum of the transitions for each stream. In our experiments, the weight of each stream was carefully adjusted to ensure the best performance can be achieved.

In our experiment, we used the hybrid DCT+LDA (64.93%) and the hybrid LBP-TOP+mRMR (54.28%) as the two streams of the MSHMM, because they achieved the best results on the VSR task using a single-stream HMM. After feeding these into multi-stream HMM using suitably adjusted weights, this yielded an accuracy of 65.24%. Next, we employed the same multi-stream HMM for DCT + LDA and LBP-TOP with our proposed cascade feature extraction scheme, and achieved a very promising result (66.49%). This indicates that our proposed scheme cannot only be used for feature-level VSR, but is also effective for classifier-level VSR. Despite the impressive accuracy from the classifier-level fusion, our proposed CHAVF achieved a better result (69.18% vs 66.49%).

Table 6: Comparison between our proposed method and the classifier-level fusion methods

Feature	Weight	Accuracy
DCT + LDA (64.93%)	0.9	65.24%
LBPTOP + mRMR (54.28%)	0.1	
DCT + LDA (64.93%)	0.8	66.49%
LBPTOP + mRMR + LDA (59.11%)	0.2	
Proposed CHAVF		**69.18%**

4.3. Speaker-Independent Visual Phrase Classification

To compare the proposed CHAVF with the state-of-the-art systems, a visual phrase classification task was performed using an SVM classifier on the popular OuluVS data corpus. In the previous continuous visual speech recognition task (see Section 4.2), both DCT and LBP features were extracted and fed into an HMM recogniser frame by frame. However, in this speech classification task, the visual features were extracted from each frame of each video, and average pooling, i.e., the mean vector of all features, was used to ensure that the visual feature has a fixed length.

Table 7 lists the classification results on the OuluVS dataset comparing the proposed method to some of the state-of-the-art systems. It shows that the proposed CHAVF was able to outperform the LBP-TOP feature [17] and sequential pattern boosting [19]. The latent variable models [24] achieved a better accuracy than the proposed method; however, the training process was more computationally complex than that of the proposed method. Furthermore, this method could not be used on the continuous visual speech recognition task. The proposed method also had an accuracy level similar to that of the transported square-root vector field [23]. However, that method was tested using a speaker-dependent condition [23] and the proposed method was tested using the more difficult speaker-independent condition.

Table 7: Visual speech classification comparison on the OuluVS data corpus. ‡ The results is reported in terms of speaker-dependent speech classification.

Method	Results
LBP-TOP (TMM2009 [17])	62.4%
Sequential Pattern Boosting (BMVC 2011 [19])	65.6%
Transported Square-Root Vector Field (CVPR 2014 [23])‡	70.6%
Latent Variable Models (PAMI2014 [24])	76.6%
Our proposed-CHAVF	**68.9**%

4.4. Speaker-Independent Audio-Visual Speech Recognition

To show that the proposed visual features could boost audio-only speech recognition, audio-visual speech recognition experiments were performed under varying noise levels. The MSHMM was used to model the audio and visual signals. The training of the MSHMM was conducted separately for the audio and visual stream under clean acoustic conditions. The test conditions were conducted under various SNR conditions representing the degradation of the audio stream.

In this study, different levels of additive white noise were used to demonstrate the robustness to audio degradation of the audio-visual speech recognition system. The proposed CHAVF and the most commonly used DCT+LDA planar (grey-level) features were used for the experiment and the corresponding audio-visual speech recognition results are listed in Figure 10.

It should be noted that the main aim of this paper was to prove the superiority of the hybrid visual features. The automatic selection of the weights for the audio and visual streams according to different noise levels was beyond the scope of this paper. Thus, the audio and visual weights (listed in Table 8) were empirically chosen to ensure that the best audio-visual fusion results were achieved.

Table 8: The Audio Weight (AW) and the Video Weight (VW) of the MSHMM.

SNR	AW	VW	SNR	AW	VW
clean	0.9	0.1	10dB	0.6	0.4
30dB	0.8	0.2	00dB	0.2	0.8
20dB	0.8	0.2	-5dB	0.1	0.9

In an acoustically clean environment (i.e., 30dB), the audio-visual fusion results were equivalent to those obtained by audio features only. However, assigning a small weight (i.e., $0.1 - 0.2$) to the visual stream audio-visual fusion results (i.e., 96.87%) still led to slightly better results than that achieved by the audio-only results (i.e., 96.52%). With an increase in the noise level, the recognition performance using audio-only features degraded significantly from 96.52% (i.e., 30dB) to 17.08% (i.e., -5dB). Conversely, the audio-visual fusion recognition performance experienced a relatively small decrease due to the utilisation of the video signals. Furthermore, it should be noted that under very noisy environments the audio-visual speech accuracy was only slightly above visual-only accuracy, as the visual modality then took on the dominant role in the recognition process.

Encouragingly, this can achieve an improvement of more than 10% with audio-visual recognition over any individual modality using the correct choice of audio and visual weights; for example, with the proposed CHAVF features, the audio-only and visual-only speech accuracy was 67.86% and 69.18%, respectively. However, the combined audio-visual system yielded an impressive accuracy (i.e., 80.49%) under the 10dB SNR condition. Thus, confirming that the complementary information provided by the visual features could be used to improve the overall recognition performance in moderate noise conditions (i.e., between 20dB and 0dB SNR).

5. Conclusion

This study investigated the integration of stereo visual features with traditional planar features to boost audio-visual speech accuracy. Notably, it was shown that the proposed novel feature extraction scheme that successfully combined planar and stereo visual information outperformed the state-of-the-art appearance features. We also showed that, even with the application

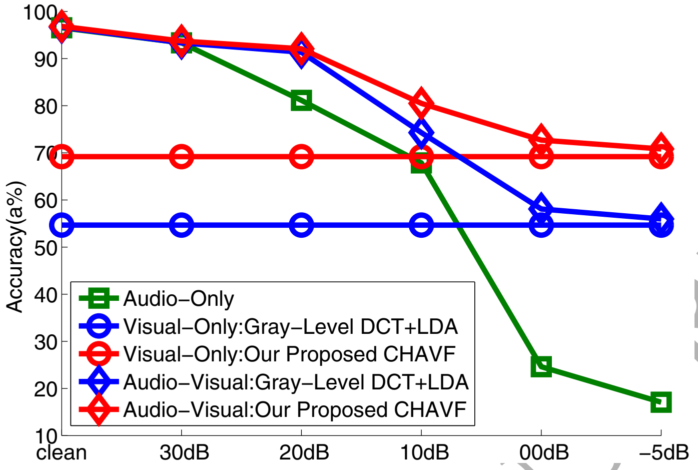


Figure 10: Multistream HMM audio-visual digit classification results with various white noise SNR levels for different types of visual features.

of low-quality stereo visual features, integrating both stereo and planar visual features led to a significant increase in visual speech accuracy.

This study also showed the different characteristics of planar and stereo features using information theoretic techniques and explained how these characteristics could benefit speech recognition. Furthermore, an analysis of the different characteristics of the planar and stereo features revealed the reasons why the stereo visual features significantly boosted the visual and audio-visual speech recognition results. After the fusion of the audio and visual signals, the experimental results showed that the proposed visual features markedly improve the audio-visual speech recognition performance in the presence of additive white noise interference.

It appears that this is the first paper to comprehensively analyse the benefits of using the hybrid (i.e., a combination of planar and stereo) visual features for the audio-visual speech recognition task on a newly collected large-scale 3D audio-visual corpus. Thus, this study provides a new perspective that effectively solves the low visual speech accuracy problem in the area of visual and audio-visual speech recognition.

Acknowledgment

The AusTalk corpus was collected as part of the Big ASC project ([38, 39, 40]), funded by the Australian Research Council (LE100100211). See:

https://austalk.edu.au/ for details.

References

[1] G. Potamianos, C. Neti, J. Luettin, I. Matthews, Issues in Visual and Audio-Visual Speech Processing, MIT Press, 2004, Ch. Audio-Visual Automatic Speech Recognition: An Overview.

[2] Y. Lei, M. Bennamoun, A. A. El-Sallam, An efficient 3D face recognition approach based on the fusion of novel local low-level features, Pattern Recognition 46 (1) (2013) 24–37.

[3] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, J. Wan, 3D object recognition in cluttered scenes with local surface features: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (11) (2014) 2270–2287.

[4] S. Sedai, M. Bennamoun, D. Q. Huynh, A Gaussian process guided particle filter for tracking 3D human pose in video, IEEE Transactions on Image Processing 22 (11) (2013) 4286–4300.

[5] G. Potamianos, C. Neti, G. Gravier, A. Garg, A. W. Senior, Recent advances in the automatic recognition of audiovisual speech, Proceedings of the IEEE 91 (9) (2003) 1306–1326.

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

[7] H. E. Cetingul, Y. Yemez, E. Erzin, A. M. Tekalp, Discriminative analysis of lip motion features for speaker identification and speech-reading, IEEE Transactions on Image Processing 15 (10) (2006) 2879–2891.

[8] M. Gurban, J.-P. Thiran, Information theoretic feature extraction for audio-visual speech recognition, IEEE Transactions on Signal Processing 57 (12) (2009) 4765–4776.

[9] V. Estellers, M. Gurban, J. Thiran, On dynamic stream weighting for audio-visual speech recognition, IEEE Transactions on Audio, Speech, and Language Processing 20 (4) (2012) 1145–1157.

[10] D. Stewart, R. Seymour, A. Pass, J. Ming, Robust audio-visual speech recognition under noisy audio-video conditions, IEEE Transactions on Cybernetics 44 (2) (2013) 175–184.

[11] G. Galatas, G. Potamianos, F. Makedon, Audio-visual speech recognition using depth information from the Kinect in noisy video conditions, in: Proceedings of International Conference on Pervasive Technologies Related to Assistive Environments, ACM, 2012, p. 2.

[12] G. Galatas, G. Potamianos, F. Makedon, Audio-visual speech recognition incorporating facial depth information captured by the Kinect, in: Proceedings of 20th European Signal Processing Conference, IEEE, 2012, pp. 2714–2717.

[13] J. Wang, Y. Gao, J. Zhang, J. Wei, J. Dang, Lipreading using profile lips rebuilt by 3D data from the Kinect, Journal of Computational Information Systems 11 (7) (2015) 2429–2438.

[14] J. Wang, J. Zhang, K. Honda, J. Wei, J. Dang, Audio-visual speech recognition integrating 3D lip information obtained from the Kinect, Multimedia Systems (2015) 1–9.

[15] X. Zhang, Y. Gao, Face recognition across pose: A review, Pattern Recognition 42 (11) (2009) 2876–2896.

[16] F. Bianconi, A. Fernández, On the occurrence probability of local binary patterns: A theoretical study, Journal of Mathematical Imaging and Vision 40 (3) (2011) 259–268.

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

[18] Z. Zhou, G. Zhao, M. Pietikainen, Towards a practical lipreading system, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 137–144.

[19] E.-J. Ong, R. Bowden, Learning sequential patterns for lipreading, in: Proceedings of 22nd British Machine Vision Conference, 2011.

[20] E.-J. Ong, R. Bowden, Learning temporal signatures for lip reading, in: Proceedings of IEEE International Conference on Computer Vision Workshops, IEEE, 2011, pp. 958–965.

[21] A. Bakry, A. Elgammal, MKPLS: Manifold kernel partial least squares for lipreading and speaker identification, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2013, pp. 684–691.

[22] Y. Pei, T.-K. Kim, H. Zha, Unsupervised random forest manifold alignment for lipreading, in: Proceedings of IEEE Conference on Computer Vision, IEEE, 2013.

[23] J. Su, A. Srivastava, F. D. de Souza, S. Sarkar, Rate-invariant analysis of trajectories on Riemannian manifolds with application in visual speech recognition, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2004.

[24] Z. Zhou, X. Hong, G. Zhao, M. Pietikäinen, A compact representation of visual speech data using latent variables, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (1) (2014) 181–187.

[25] I. Matthews, G. Potamianos, C. Neti, J. Luettin, A comparison of model and transform-based visual features for audio-visual LVCSR, in: Proceedings of IEEE International Conference on Multimedia and Expo, IEEE, 2001, pp. 825–828.

[26] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, R. Harvey, Extraction of visual features for lipreading, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2) (2002) 198–213.

[27] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Ng, Multimodal deep learning, in: Proceedings of 28th International Conference on Machine Learning, 2011, pp. 689–696.

[28] G. Potamianos, P. Scanlon, Exploiting lower face symmetry in appearance-based automatic speechreading, in: Proceedings of International Conference on Auditory-Visual Speech Processing, 2005, pp. 79–84.

[29] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (7) (2002) 971–987.

[30] B. Scholkopft, K.-R. Mullert, Fisher discriminant analysis with kernels, in: Proceedings of IEEE Signal Processing Society Workshop Neural Networks for Signal Processing, 1999, pp. 23–25.

[31] H. Yu, J. Yang, A direct LDA algorithm for high-dimensional datawith application to face recognition, Pattern recognition 34 (10) (2001) 2067–2070.

[32] R. Fano, Transmission of Information: A Statistical Theory of Communications, M.I.T. Press, 1961.

[33] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (8) (2005) 1226–1238.

[34] M. Vidal-Naquet, S. Ullman, Object recognition with informative features and linear classification, in: Proceedings of IEEE International Conference on Computer Vision, IEEE, 2003, pp. 281–288.

[35] A. Ortega, F. Sukno, E. Lleida, R. Frangi, A. Miguel, L. Buera, E. Zacur, AV@CAR: A Spanish multichannel multimodal corpus for in-vehicle automatic audio-visual speech recognition, in: Proceedings of Inernational Conference on Language Resources and Evaluation, 2004, pp. 763–767.

[36] A. Vorwerk, X. Wang, D. Kolossa, S. Zeiler, R. Orglmeister, WAPUSK20 - a database for robust audiovisual speech recognition, in: Proceedings of the 7th International Conference on Language Resources and Evaluation, European Language Resources Association (ELRA), 2010.

[37] R. Goecke, J. B. Millar, The audio-video Australian English speech data corpus AVOZES, in: Proceedings of 10th Australian International Conference on Speech Science and Technology, 2004, pp. 486–491.

[38] D. Burnham, E. Ambikairajah, J. Arciuli, M. Bennamoun, C. T. Best, S. Bird, A. Butcher, C. Cassidy, G. Chetty, F. M. Cox, et al., A blueprint

for a comprehensive Australian English auditory-visual speech corpus, in: Proceedings of HCSNet Workshop on Designing the Australian National Corpus, 2009, pp. 96–107.

[39] M. Wagner, D. Tran, R. Togneri, P. Rose, D. Powers, M. Onslow, D. Loakes, T. Lewis, T. Kuratate, Y. Kinoshita, et al., The big Australian speech corpus (the big ASC), in: Proceedings of 13th Australasian International Conference on Speech Science and Technology, 2010, pp. 166–170.

[40] D. Burnham, D. Estival, S. Fazio, J. Viethen, F. Cox, R. Dale, S. Cassidy, J. Epps, R. Togneri, M. Wagner, et al., Building an audio-visual corpus of Australian English: Large corpus collection with an economical portable and replicable black box, in: Proceedings of Annual Conference of the International Speech Communication Association, 2011.

[41] C. Sui, S. Haque, R. Togneri, M. Bennamoun, A 3D audio-visual corpus for speech recognition, in: Proceedings of Australasian International Conference on Speech Science and Technology, 2012.

[42] 3dMD — The world leader in anatomically-precise 3D and "temporal-3D" (4D) suface imaging systems and software, http://www.3dmd.com/ (2016 (accessed October 1, 2016)).

[43] P. Viola, M. J. Jones, Robust real-time face detection, International Journal of Computer Vision 57 (2) (2004) 137–154.

[44] P. J. Besl, N. D. McKay, A method for registration 3D shapes, IEEE Transactions on Pattern Analysis and Machine Intelligence 14 (2) (1992) 239–256.

[45] A. S. Mian, M. Bennamoun, R. Owens, An efficient multimodal 2D-3D hybrid approach to automatic face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (11) (2007) 1927–1943.

[46] E. K. Patterson, S. Gurbuz, Z. Tufekci, J. N. Gowdy, CUAVE: A new audio-visual database for multimodal human-computer interface research, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, IEEE, 2002, pp. 2017–2020.

[47] S. J. Young, G. Evermann, M. Gales, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, The HTK book version 3.5, Cambridge University Engineering Department, 2015.

[48] C. Sui, R. Togneri, S. Haque, M. Bennamoun, Discrimination comparison between audio and visual features, in: Proceedings of 46th Asilomar Conference on Signals, Systems and Computers, IEEE, 2012, pp. 1609–1612.

[49] H. Hotelling, Relations between two sets of variates, Biometrika 28 (3/4) (1936) 321–377.

[50] H. Gao, J. W. Davis, Why direct LDA is not equivalent to LDA, Pattern Recognition 39 (5) (2006) 1002–1006.

[51] S. Sun, A survey of multi-view machine learning, Neural Computing and Applications 23 (7) (2013) 2031–2038.

[52] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, Journal of Machine Learning Research 9 (2008) 2579–2605.