Visual Speech Feature Representations: Recent Advances

Chao Sui, Mohammed Bennamoun, and Roberto Togneri

Abstract Exploiting the relevant speech information that is embedded in facial images has been a significant research topic in recent years, because it has provided complementary information to acoustic signals for a wide range of automatic speech recognition (ASR) tasks. Visual information is particularly important in many real applications where acoustic signals are corrupted by environmental noises. This chapter reviews the most recent advances in feature extraction and representation for Visual Speech Recognition (VSR). In comparison with other surveys published in the past decade, this chapter presents a more up-to-date survey and highlights the strengths of two newly developed approaches (i.e., graph-based learning and deep learning) for VSR. In particular, we summarise the methods of using these two techniques to overcome one of the most challenging difficulties in this area-that is, how to automatically learn good visual feature representations from facial images to replace the widely used handcrafted features. This chapter concludes by discussing potential visual feature representation solutions that may overcome the remaining challenges in this domain.

1 Introduction

Given that speech is widely acknowledged to be one of the most effective means of communication between humans, researchers in the automatic speech recognition (ASR) community have made great efforts to provide users with a natural way to communicate using intelligent devices. This is particularly important for disabled people, who may be incapable of using a keyboard, mouse or joystick. As a result of the great achievements made by the ASR community in recent years in terms of the

C. Sui (✉) • M. Bennamoun
School of Computer Science and Software Engineering, University of Western Australia, Perth, WA, Australia
e-mail: chao.sui@uwa.edu.au; mohammed.bennamoun@uwa.edu.au

R. Togneri
School of Electrical, Electronic and Computer Engineering, University of Western Australia, Perth, WA, Australia
e-mail: roberto.togneri@uwa.edu.au

M. Kawulok et al. (eds.), *Advances in Face Detection and Facial Image Analysis*,
DOI 10.1007/978-3-319-25958-1_14

Fig. 1 Possible application scenarios of VSR. In an acoustically noisy environment, using an intelligent handset to capture and extract visual features is an effective solution for ASR

application of novel techniques such as deep learning [26], people generally believe that we are getting closer to talking naturally and freely to our computers[12].

Although a number of ASR systems have been commercialised and have entered our daily lives (e.g., Apples Siri and Microsofts Cortana), several limitations still exist in this area. One major limitation is that ASR systems are still prone to environmental noises, thereby limiting their applications. Given ASR's vulnerability, research in the area of Visual Speech Recognition (VSR) has emerged to provide an alternative solution to improve speech recognition performance. Further, VSR systems have a wider range of applications compared to their acoustic-only speech recognition counterparts. For example, as shown in Fig. 1, in many practical applications where speech recognition systems are exposed to noisy environments, acoustic signals are almost unusable for speech recognition. Conversely, with the availability of front and rear cameras on most intelligent mobile devices, users can easily record facial movements to perform VSR. In extremely noisy environments, visual information basically becomes the only source that ASR systems can use for speech recognition.

Moreover, inspired by bimodal human speech production and perception even in clean and moderate noise conditions, where good-quality acoustic signals are available for speech recognition visual information can provide complementary information for ASR [54, 55]. Therefore, research on VSR is of particular importance, because once an adequate VSR result is obtained, speech recognition performance can be boosted through the fusion of audio and visual modalities.

Despite the wide range of applications of VSR systems, there are two main limitations related to this area: the development of appropriate dynamic audio-visual fusion and the development of appropriate visual feature representations. Regarding dynamic audio-visual fusion, although several high-quality works on this topic have been published recently [15, 49, 61, 64], a similar fusion framework was used in most of the cases. More specifically, in these works, the quality of both the audio and visual signals was evaluated using different criteria, such as signal-to-noise ratio, dispersion and entropy. Weights were dynamically assigned to the audio and visual

streams according to the quality of the audio and visual signals. However, compared with the audio-visual fusion method, visual feature representation techniques are more controversial. The goal of visual feature representation is to embed spatio-temporal visual information into a compact visual feature vector. This is the most fundamental problem for VSR, because it directly affects the final recognition performance. Hence, in this survey, we mainly focus on the most recent advances in the area of visual feature representation, and we discuss potential solutions and future research directions.

Regarding audio feature representation, Mel-Frequency Cepstral Coefficients (MFCCs) are generally acknowledged to be the most widely used acoustic features for speech recognition. However, unlike audio feature extraction, there is no universally accepted visual feature extraction technique that can achieve promising results for different speakers and different speech tasks, as three fundamental issues remain unresolved [79]: (1) how to extract visual features with constant quality from videos with different head and pose positions; (2) how to remove speech-irrelevant information from the visual data; (3) how to encode temporal information into the visual features. This chapter will summarise recent research that has examined solutions to these issues, and it will provide an insight into the relationships between these methods.

This chapter is organised as follows. Section 2 introduces handcrafted visual feature extraction methods, which are still the most widely used techniques for visual feature representation, and they are sometimes used in pre-processing steps for automatic feature learning. Sections 3 and 4 respectively describe graph-based feature learning and deep learning-based feature learning methods. Finally, Sect. 5 provides insights into potential solutions for the remaining challenges and possible future research directions in this area.

2 Hand Crafted Feature Extraction

Before introducing visual feature learning techniques, this section describes some of the handcrafted visual features that still play a dominant role in VSR. In addition, handcrafted feature extraction methods can be used in the pre-processing steps of many visual feature learning frameworks. In terms of the type of information embedded in the features, visual features can be categorised into two classes: appearance-based and geometric-based features [7].

For appearance-based visual features, the entire ROIs (e.g., mouth, lower face or even the face area) are considered informative regions in terms of VSR. However, it is infeasible to use all the pixels of ROIs because the dimensions of the features are too large for the classifiers to process. Hence, appropriate transformations of the ROIs are used to map the images to a much lower-dimensional feature space. More specifically, given the original image $\mathbf{I}$ in the feature space $\mathbb{R}^D$ (where D is the feature dimension), appearance-based feature extraction methods seek to transform

matrix $\mathbf{P}$ to map $\mathbf{I}$ to a lower feature space $\mathbb{R}^d$ ($d << D$), such that the transformed feature vector contains the most speech-relevant information with a much smaller feature dimension.

The Discrete Cosine Transform (DCT) [54, 55] is among the most commonly used appearance-based visual feature extraction methods. It can be formulated as:

$$Y(i,j) = \sum_{j=1}^{N} \sum_{i=1}^{N} I(i,j) \cos\left(\frac{\pi(2j+1)j}{2N}\right) \cos\left(\frac{\pi(2i+1)i}{2N}\right), \quad (1)$$

for $i,j = 1,2,\ldots,N$, where N is the width and height of the mouth ROI, the value of N is a power of two and $I(i,j)$ is the grey-level intensity value of the ROI. To avoid the curse of dimensionality, low-frequency coefficients are selected and used as the static components of the visual feature. To encode the temporal information, the first and second derivatives of the DCT coefficients are used along with the static coefficients of the DCT ($Y(i,j)$) as the dynamic components of the visual feature. Other appearance-based techniques can also be used to extract appearance-based visual features [55], such as Principle Component Analysis (PCA) [13], Hadamard and Haar transform [60] and Discrete Wavelet Transform (DWT) [53].

In addition to the methods described above, other appearance-based visual feature extraction methods have been proposed. More specifically, instead of seeking a global transformation that can be used on the entire ROI, other methods use a feature descriptor to describe a small region centred at each pixel in the ROI, and to count the descriptors response occurrence in the ROI. Typical methods in this category include Local Binary Pattern (LBP) [45] and Histogram of Oriented Gradients (HOG) [9]. However, these methods are incapable of extracting temporal dynamic information from the ROIs. Hence, a number of variants have been proposed. For example, Zhao et al. [74] proposed a local spatio-temporal visual feature descriptor for automatic lipreading. This visual feature descriptor can be viewed as an extension of the basic LBP [45]. More specifically, to encode the temporal information into the visual feature vector, Zhao et al. [74] extracted LBP features from Three Orthogonal Planes (LBP-TOP), which contain the spatial axes of the images (X and Y) and the time axis (T), as shown in Fig. 2. Although the LBP-TOP feature contains rich visual speech-relevant information, the dimensionality of the original LBP-TOP feature is too large to be used directly for VSR. Hence, in [74], AdaBoost was used to select the most informative components from the original LBP-TOP feature for VSR.

Numerous works [2, 77, 78] have used LBP-TOP for VSR, and variations of the original LBP-TOP feature have also been proposed. Pei et al. [51] used Active Appearance Models (AAM) to track keypoints on the lips. For each small patch centred around the keypoints of the lips, LBP-TOP and HOG were used to extract the texture features. In addition to the texture features, the difference between the patch positions in the adjacent frames was used as a shape feature. Given that rich speech-relevant information is embedded in LBP-TOP features, a number of feature reduction techniques have been introduced to extract a more compact visual feature

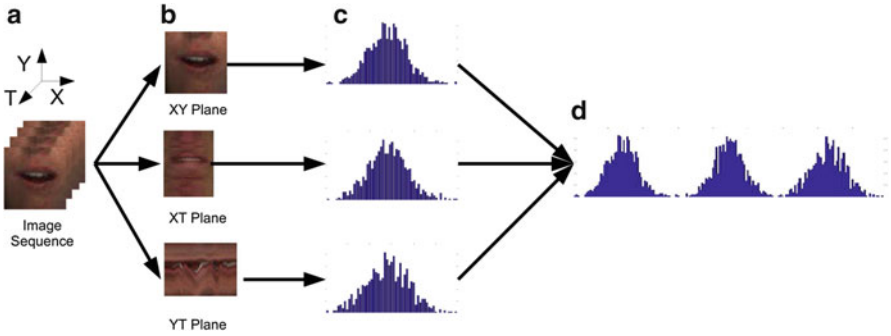

Fig. 2 Lip spatio-temporal feature extraction using LBP-TOP feature extraction. (**a**) Lip block volumes; (**b**) lip images from three orthogonal planes; (**c**) LBP features from three orthogonal planes; (**d**) concatenated features for one block volume with appearance and motion

from the original LBP-TOP feature. In addition to the LBP-TOP feature, a number of other appearance-based feature descriptors have been used to extract temporal dynamic information, such as LPQ-TOP [30], LBP-HF [76], and LGBP-TOP [1].

Although both DCT and LBP-TOP are widely used for VSR, they are quite different because they represent visual information from different perspectives. More specifically, as shown in (1), each component ($Y(i,j)$) of the DCT feature is a representation of the entire mouth region at a particular frequency. Hence, 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 method. Hence, the development of a method that can combine both global and local information using a compact feature vector would be expected to boost visual speech accuracy. Although Zhao et al. [75] showed that combining different types of visual features (LBP-TOP and EdgeMap [17]) can improve recognition accuracy, finding an effective way to combine DCT and LBP-TOP features is still an undeveloped area.

Although the dimensionality of the appearance-based visual features is much smaller compared to the number of pixels in the ROI, it still makes the system succumb to the curse of dimensionality. Hence, a feature dimension reduction process is essential as a prior step to VSR. Among the feature reduction methods, LDA and PCA are the most widely used [54, 55]. In addition, Gurban et al. [20] presented a Mutual Information Feature Selector (MIFS)-based scheme to select an informative visual feature component subset and thus reduce the dimensionality of the visual feature vector. Unlike feature reduction schemes such as PCA and LDA, MIFS analyses each feature component in the visual feature vector and selects the most informative components using the greedy algorithm. In addition, Gurban et al. [20] proposed that penalizing features for their redundancy is essential to yield a more informative visual feature vector.

Geometric visual features explicitly model the shape of the mouth and are potentially more powerful than appearance-based features. However, they are sensitive to lighting conditions and image quality. Geometric-based features include Deformable Template (DT), Active Shape Model (ASM), Active Appearance Model (AAM) and Active Contour Model (ACM). DT [36] is a method that uses a parametric lip template to partition an input image into a lip region and a non-lip region. However, this approach is degraded when the shape of the lip is irregular or the mouth is opened wide [31]. The ASM [42] uses a set of landmarks to describe the lip model. The AAM approach [38] can be viewed as an extension of ASM that incorporates grey-level information into the model. However, as the landmarks need to be manually labelled during training, it is very laborious and time-consuming to train the ASM and AAM for lip extraction.

In terms of ACM-based lip extraction, there are two main categories, namely edge-based and region-based. With respect to the edge-based extraction approach, the image gradients are calculated to locate the lip potential boundary [11]. Unfortunately, given that the intensity contrast between the lip and the face region is usually not large enough, the edge-based ACM is likely to achieve incorrect extraction results. Moreover, this method has been confirmed to be prone to image noise, and it is highly dependent on the initial parameters of the ACM [31]. In terms of region-based techniques, the foreground is segmented from the background by finding the optimum intensity energy in the images. Compared to its edge-based counterpart, this method has been shown to be robust with respect to the initial curve selection and the influence of noise [31]. In contrast, because of the appearance of the teeth and tongue, intensity values inside the lips are usually different. In this situation, a Global region-based ACM (GACM) can fail because all of the pixels inside the lips are taken into consideration. However, with a Localised region-based ACM (LACM), only the pixels around the objects contour are taken into account. This method can therefore successfully avoid the influence of the appearance of the teeth and the tongue [8].

However, provided that the LACM is used solely for lip extraction and the initial contour is far away from the actual lip contour, the curvature may converge to a local minima without finding the correct lip boundary. Therefore, the initial contour needs to be specified near the lip boundary as a priori. The common method for specifying the initial contour is to detect several lip corners [10, 35] and to construct an ellipse surrounding the lip. Unfortunately, this approach is either sensitive to the image noise and illuminations or needs a complex training process. In order to effectively solve this problem, Sui et al. [69] presented a new extraction framework that synthesises the advantages of both the global and localised region-based ACMs.

Although the geometric feature can explicitly model the shape of the lips, it is difficult to derive an accurate model that can describe the dynamic movement of the mouth. Hence, appearance-based features remain the most widely used features in the VSR community.

3 Graph Based Visual Feature Representations

In most cases, the dimensions of the visual features that are extracted using handcrafted feature extraction methods are usually too large for the classifiers. Graph-based learning methods that non-linearly map the original visual features to a more compact and discriminatory feature space have also been used in recent years.

Initially, graph-based methods were commonly used in human activity recognition [52]. Given that both human activity and speech recognition deal with the analysis of spatial and temporal information, graph-based feature learning can therefore also be used for VSR. The idea behind graph-based learning is that visual features can be represented as the elements of a unified feature space, and the temporal evolution of lip movements can be viewed as the trajectory connecting these elements in the feature space. Hence, after the extracted feature sequences from the videos have been correctly mapped to the corresponding trajectories, the speech can be correctly recognised. In addition, it is generally believed that the dimension of the underlying structure of the visual speech information should be significantly smaller than the dimension of the corresponding observed videos. Based on the above assumptions, numerous papers have proposed different frameworks to parameterise the original high-dimensional visual features to the trajectories to extract lower-dimensional features. An illustration of the concept behind graph-based feature representation methods is shown in Fig. 3.

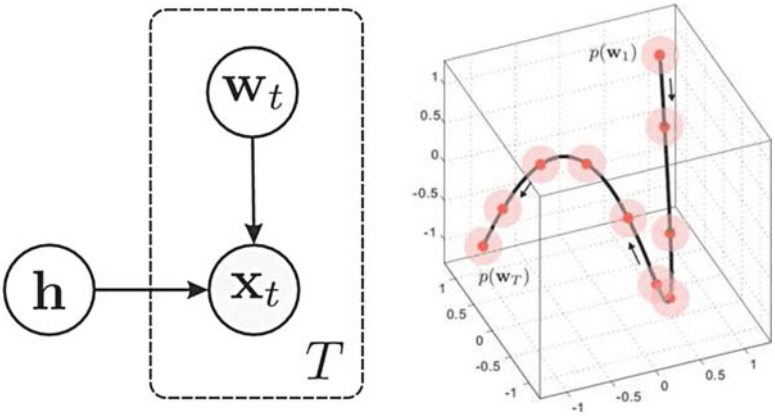

Fig. 3 The idea behind graph-based feature representation methods is to project the original high-dimensional spatio-temporal visual features to a trajectory in a lower-dimensional feature space, thereby reducing the feature dimension to boost the performance of speech recognition. Each point $(p(\mathbf{w}_T))$ of the projected trajectory represents a frame in the corresponding video. This figure appeared in [78]. In this work, each image $\mathbf{x}_i$ of the T-frame video is assumed to be generated by the latent speaker variable h and the latent utterance variable $\mathbf{w}_i$

Zhou et al. [77] proposed a path graph based method to map the image sequence of a given utterance to a low-dimensional curve. Their experimental results showed that the recognition rate of this method is 20 % higher than the recognition rate reported in [74] on the OuluVS data corpus. Based on this work, the visual feature sequence of a speakers mouth when talking is further assumed to be generated from a speaker-dependent Latent Speaker Variable (LSV) and a sequence of speaker-independent Latent Utterance Variables (LUV). Hence, Zhou et al. [78] presented a Latent Variable Model (LVM) that separately represents the video by LSV and LUV, and the LUV is further used for VSR. Given an image sequence of length T, $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^{T}$, the LVM of an image $\mathbf{x}_t$ which is generated from the inter-speaker variations h (LSV) and dynamic changes of the mouth $\mathbf{w}_t$, can be formulated by (2):

$$\mathbf{x}_t = \boldsymbol{\mu} + \mathbf{F}\mathbf{h} + \mathbf{G}\mathbf{w}_t + \boldsymbol{\epsilon}_t, \tag{2}$$

where μ is the global mean, $\mathbf{F}$ is a factor matrix whose columns span the inter-speaker space, $\mathbf{G}$ is the bias matrix that describes the uttering variations and $\boldsymbol{\epsilon}_t$ is the noise term. The model described in (2) is a compact representation of high-dimensional visual features. Compared with the 885-dimensional raw LBP-TOP feature, the six-dimensional LUV feature is very compact and can yield better accuracy than other features, such as PCA [4], DCT [18], AF[57] and AAM [38].

Pei et al. [51] presented a method based on the concept of unsupervised random forest manifold alignment. In this work, both appearance and geometric visual features were extracted from the lip videos, and the affinity of the patch trajectories in the lip videos was estimated by a density random forest. A multidimensional scaling algorithm was then used to embed the original data into a low-dimensional feature space. Their experimental results showed that this method was capable of handling large datasets and low-resolution videos effectively. Moreover, the exploitation of depth information for VSR was also discussed in this paper.

Unlike the unsupervised manifold alignment approach proposed by Pei et al. [51], Bakry and Elgammal [2] presented a supervised visual feature learning framework where each video was first mapped to a manifold by the manifold parametrisation [14], and then kernel partial least squares was used in the manifold parameterisation space to yield a latent low-dimensional manifold parameterisation space.

It is well known that different people speak at different rates, even when they are uttering the same word. The varying rates of speech result in random parameterisations of the same trajectory, which leads to a failure in speech recognition. Hence, a temporal alignment is essential for VSR to remove any temporal variabilities caused by different speech rates. Su et al. [65] applied a statistical framework (introduced in [66]) and proposed a rate-invariant manifold alignment method for VSR. In this method, each trajectory α of the video sequence in the trajectory set $\mathbb{M}$ is represented by a Transported Square-Root Vector Field (TSRVF) to a reference point c:

$$h_\alpha(t) = \frac{\dot{\alpha}(t)_{\alpha(t)\to c}}{\sqrt{|\dot{\alpha}(t)|}}, \tag{3}$$

where $h_\alpha(t)$ is the TSRVF of trajectory α at time t, $\dot{\alpha}(t)$ it the velocity vector of $\alpha(t)$, and $|\cdot|$ is defined as the Riemannian metric on the Riemannian manifold. Given the TSRVFs of two smooth trajectories α_1 and α_2, these two trajectories can be aligned according to:

$$\gamma^* = \arg\min_{\gamma\in\Gamma} \sqrt{\int_0^1 |h_{\alpha_1}(t) - h_{\alpha_2}(\gamma(t))\sqrt{\dot{\gamma}(t)}|^2 \mathrm{d}t}, \tag{4}$$

where Γ is the set of all diffeomorphisms of $[0, 1] : \Gamma = \{\gamma : [0, 1] \to [0, 1] | \gamma(0) = 0, \gamma(1) = 1, \gamma$ is a diffeomorphism$\}$. The minimization over Γ in (4) can be solved using dynamic programming. After the trajectories have been registered, the mean of the multiple trajectories can be used as a template for visual speech classification. Although the method introduced in [65] did not produce superior performance over other recent graph-based methods [2, 51, 77, 78], and although only speech-dependent recognition was reported, this work provided a general mathematical speech-rate-invariant framework for the registration of trajectories and for comparison.

Despite graph-based methods have shown promising recognition performance compared to conventional feature reduction methods [79] such as LDA and PCA, it should be noted that none of the above graph-based methods were tested on continuous speech recognition. Even though Zhou et al. [78] reported that their method achieved promising results on classifying visemes, which are generic images that can be used to describe a particular sound, it is still unclear whether their graph-based method can be used for continuous speech recognition.

4 Visual Feature Learning Using Deep Learning

Section 3 introduced various graph-based methods that can map high-dimensional visual features to non-linear feature spaces. However, the use of graph-based methods for VSR requires prior extraction of the visual features, and the classification performance largely depends on the quality of the extracted visual features. In this section, we introduce deep feature learning-based methods, which can directly learn visual features from videos. These techniques offer the potential to replace handcrafted features with deep learned features for the VSR task.

Deep learning techniques were first proposed by Hinton et al. [25], who used the greedy, unsupervised, layer-wise pre-training scheme to train a Restricted Boltzmann Machine (RBM) to model each layer of a Deep Belief Network (DBN), which effectively solved the difficulty of training multiple hidden-layer neural networks. Later works showed that a similar pre-training scheme could also be

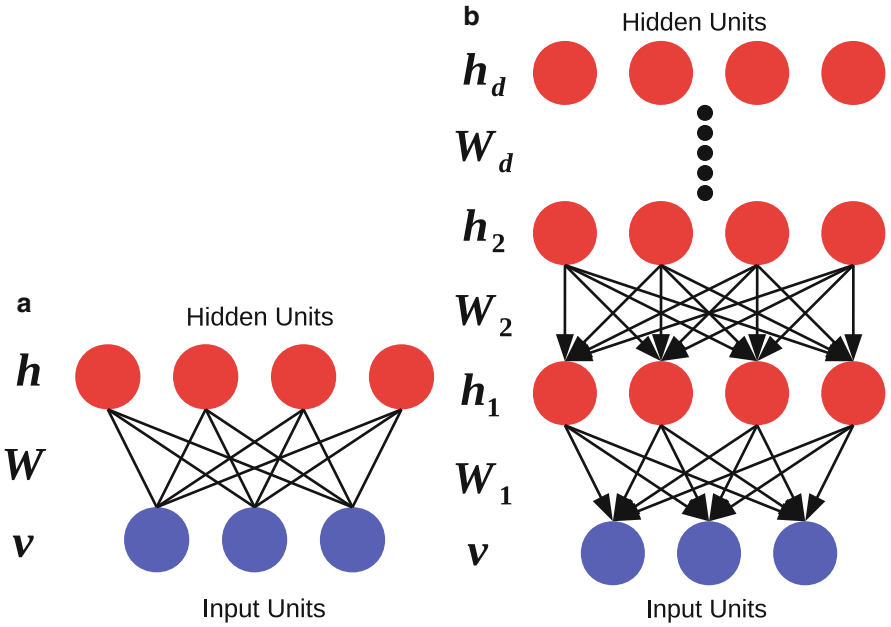

Fig. 4 Two RBM-based deep models. *Blue* circles represent input units and *red* units represent hidden units. (**a**): An RBM. (**b**): A Stacked RBM-based Auto-Encoder

used by stacked auto-encoders [3] and Convolutional Neural Networks (CNN) [56]. These techniques achieved great success in various classification tasks, such as acoustic speech recognition and image set classification [21, 22].

After deep learning techniques had been successfully applied to a single modality for the task of feature learning, Ngiam et al. [41] used it for a bimodal (i.e., audio and video) task. This was the first deep learning work in the domain of VSR and Audio-Visual Speech Recognition (AVSR). Since then, a number of other methods have been proposed that employed deep learning techniques to learn visual features for visual speech classification. Deep learning techniques used for VSR and AVSR can be categorised into three types: RBM-based deep models, stacked denoising auto-encoder-based methods and CNN-based methods.

The RBM is a particular type of Markov random field with hidden variables $\mathbf{h}$ and visible variables $\mathbf{v}$ (Fig. 4a). The connections W_{ij} between the visible and hidden variables are symmetrical, but there are no connections within the hidden and visible variables. The model defines the probability distribution $P(\mathbf{v}, \mathbf{h})$ over $\mathbf{v}$ and $\mathbf{h}$ via an energy function, which can be formulated by (5). The log-likelihood of $P(\mathbf{v}, \mathbf{h})$ can be maximised by minimising the energy function in (5):

$$E(\mathbf{v}, \mathbf{h}; \theta) = -\sum_{i=1}^{m}\sum_{j=1}^{n} W_{ij} v_i h_j - \sum_{i=1}^{m} b_i v_i - \sum_{j=1}^{n} a_j h_j, \tag{5}$$

where a_j and b_i are the biases of the hidden units and visible unit respectively, m and n are the numbers of hidden units and visible units, and θ includes the parameters of the model. As the computation of the gradient of the log-likelihood is intractable, the parameters of the model have usually been learned using contrastive divergence [24]. With the proper configurations of the RBM, the visual feature is fed to the first layer of the RBM, the posteriors of the hidden variables (given the visible variables) are obtained using $p(h_j|\mathbf{v}) = \text{sigmoid}(b_j + W_j^T\mathbf{v})$, and $p(h_j|\mathbf{v})$ can be used as the new training data for the successive layers of the RBM-based deep networks. This process is repeated until the subsequent layers are all pre-trained.

In Ngiam et al.'s work [41], the deep auto-encoder, which consisted of multiple layers of sparsity RBMs [34], was used to learn a shared representation of the audio and visual modalities for speech recognition. The authors discussed two learning architectures in their paper. The first model investigated was cross-modality learning, where the model learned to reconstruct both the audio and video modalities, while only the video was used as an input during the training and testing stage. The second model was used for the training of the multimodal deep auto-encoder with both audio and video data. However, two-thirds of the used data had zero values in one of the input modalities (e.g., video), and the original values were used in the other input modality (e.g., audio). Experimental results in [41] showed an improvement over previous handcrafted visual features [20, 38, 74]. However, their bimodal deep auto-encoder did not outperform their video-only deep auto-encoder, because the bimodal auto-encoder might not have been optimal when only the visual input was provided.

Given the inefficiency of the bimodal auto-encoders proposed in [41], Srivastava et al. [62] used a Deep Boltzmann Machine (DBM), which was first proposed in [59], for AVSR. Like the deep learning models introduced above, the DBM is also a method from the Boltzmann machine family of models, and it has the potential to learn the complex and non-linear representations of the data. Moreover, it can also exploit information from a large amount of unlabelled data for pre-training purposes. The major difference between the DBM and other RBM-based models is shown in Fig. 5. Unlike other RBM-based models, which only employ a top-down approximation inference procedure, the DBM incorporates a bottom-up pass with a top-down feedback. Given that the approximation inference procedure of the DBM has two directions, the DBM model is an undirected model (Fig. 5b), while other RBM-based models are directed (Figs. 4b and 5a). Because of the undirected characteristics of the DBM models, the DBM is more capable of handling uncertainty in the data, and it is more robust to ambiguous inputs [59].

Before applying the DBM model to AVSR, Srivastava et al. [63] first applied the DBM on image and text classification, which is also a multimodal learning task. In their work, the image and text data were trained separately using two single-stream DBMs, and the outputs of these two single-stream DBMs were then merged to train joint representations of the image and text data. As the image and text data are highly correlated, it is difficult for the model proposed in [41] to learn these correlations and produce multimodal representations. In fact, as the approximation inference procedure is directed, the responsibility of the multimodal modelling falls

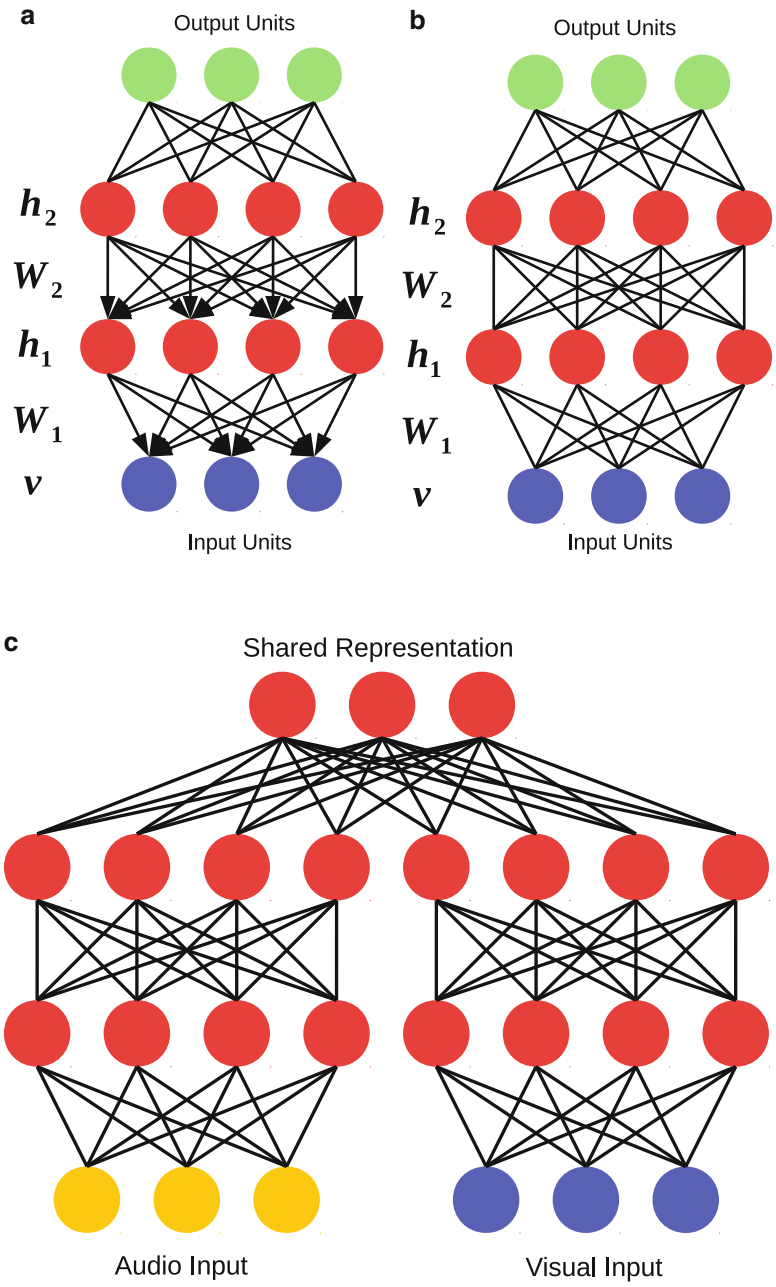

Fig. 5 Different deep models. The *blue* and *orange* circles represent input units, the *red* units represent hidden units, and the *green* circles represent representation units. (**a**): A DBN. (**b**): A DBM. (**c**): A multimodal DBM. When we compare (**a**) with (**b**), one can note that the DBN model is a directed model, while the DBM model is undirected

entirely on the joint layer [63]. In contrast, the model introduced in [63] solved this challenge effectively because the DBM can approximate the learning model both from the top-up pass and the bottom-down feedback, which makes the multimodal modelling responsibility spread out over the entire network [63].

Moreover, as shown in Fig. 5a, the top two layers of the DBN consist of an RBM (which is an undirected model), while the remaining lower layers form a directed generative model. Hence, the directed DBN model is not capable of modelling the missing inputs. Conversely, as the DBM is an undirected generative model and employs a two-way approximate inference procedure, it can be used to generate a missing modality by clamping the observed modality at the inputs and running the standard Gibbs sampler. In [63], the DBM was shown to be capable of generating missing text tags from the corresponding images. Srivastava et al. [62] then used this model for the task of AVSR. Experimental results on the CUAVE [50] and AVLetters [38] datasets showed that the multimodal DBM can effectively combine features across modalities and achieve slightly better results than the video deep auto-encoder proposed in [41]. Although this work demonstrated that the DBM could combine features effectively for speech recognition across audio and visual modalities, the inference of audio from the visual feature was not discussed. However, it provides a method that may be able to solve the problem proposed in [41]-that is, how to generate the missing audio from the video.

Despite these promising results, it should be noted that all of the aforementioned deep learning-based VSR methods have the objective of learning a more informative spatio-temporal descriptor that extracts speech-relevant information directly from the video. However, in order to use deep learning techniques for real-world VSR applications, sequential inference-based approaches, which are widely used by the acoustic speech recognition community, need to be developed.

In terms of acoustic continuous speech recognition, Mohamed et al. [40] developed acoustic phone recognition using a DBN. In this work, MFCCs were used as an input to the DBN. The DBN was pre-trained layer by layer, followed by a fine-tuning process that used 183 target class labels (i.e., three states for each of the 61 phonemes). The output of the DBN represents the probability distribution over possible classes. The probability distribution yielded by the DBN was fed to a Viterbi decoder to generate the final phone recognition results. Inspired by this method, Huang and Kingsbury [27] presented a similar framework for AVSR. Compared with the Hidden Markov Model/Gaussian Mixture Model (HMM/GMM) framework, the DBN achieved a 7 % relative improvement on the audio-visual continuously spoken digit recognition task. This work also presented a mid-level feature fusion method that concatenated the hidden representations from the audio and visual DBN, and the LDA was then used to reduce the dimensionality of the original concatenated hidden representations. At the last stage, the LDA projected representations were used as inputs to train a HMM/GMM model, and achieved a 21 % relative gain over the baseline system. However, using the DBN for visual-only speech recognition did not produce any improvements over the standard HMM/GMM model in [27].

In addition to the RBM-based deep learning techniques introduced above, Vincent et al. [71] proposed a Stacked Denoising Auto-encoder (SDA) based on a new scheme to pre-train a multi-layer neural network. Instead of training the RBM to initialise the hidden units, the hidden units are learned by reconstructing input data from artificial corruption. Paleček [47] explored the possibility of using the auto-encoder to learn useful feature representations for the VSR task. The learned features were further processed by a hierarchical LDA to capture the speech dynamics before feeding them into the HMM for classification. The auto-encoder-learned features produced a 4–8 % improvement in accuracy over the standard DCT feature in the case of isolated word recognition. However, only the single-layer auto-encoder was discussed in their paper [47], suggesting that the superiority of the stacked auto-encoder was not fully analysed. In addition to the conventional SDA, deep bottleneck feature extraction methods based on SDA [16, 58, 73] were extensively used in acoustic speech recognition. Inspired by the deep bottleneck audio features for continuous speech recognition, Sui et al. [70] developed a deep bottleneck feature learning scheme for VSR. This technique was successfully used with the connected word VSR, and it demonstrated superior performance over handcrafted features such as DCT and LBP-TOP [70].

Although RBM-based deep networks and SDA-based methods achieved an impressive performance for various tasks, these techniques did not take the topological structure of the input data into account (e.g., the 2D layout of images and the 3D structure of videos). However, topological information is very important for visual-driven tasks, because a large amount of speech-relevant information is embedded in the topological structure of the video data. Hence, developing a method to explore the topological structure of the input should help to boost VSR performance. The CNN model proposed by Lecun et al. [32] can exploit the spatial correlation that is presented in input images. This model has achieved great success in visual-driven tasks in recent years [33]. Noda et al. [43, 44] developed a lipreading system based on a CNN to recognise isolated Japanese words. In their paper, the CNN was trained using mouth images as input to recognise the phonemes. The parameters of the fully trained CNN were used as features for the HMM/GMM models. The experimental results showed that their proposed CNN-learned features significantly outperformed those acquired by PCA.

A number of deep learning-based methods have achieved promising results in the case of acoustic speech recognition. However, their use in the task of VSR has not yet been explored. For example, deep recurrent neural networks [19] have been recently proposed for acoustic speech recognition. It would be interesting to explore their applications to VSR in future research.

5 Discussion

This chapter provides an overview of some handcrafted, graph-based and deep learning-based visual features that have recently been proposed. To compare the VSR performance achieved by the different visual feature representations, we

Table 1 Summary of the recently proposed multi-speaker and speaker-independent visual-only speech recognition performance on popular and publicly available visual speech corpora

Data corpus	Feature category	Feature extraction methods	Classifier	Accuracy (%)
AVLetters	Hand crafted	ASM [38]	HMM	26.91
		Optical Flow	SVM	32.31
		AAM [38]	HMM	41.9
		MSA [38]	HMM	44.6
		DCT	SVM	53.46
		LBP-TOP [74]	SVM	58.85
	Graph-based	Bakry and Elgammal [2]	SVM	65.64
	Deep learning	Ngiam et al. [41]	SVM	64.4
		Srivastava et al. [62]	SVM	64.7
OuluVS	Hand crafted	LBP-TOP [74]	SVM	62.4
	Graph-based	Ong and Bowden [46]	SVM	65.6
		Zhou et al. [77]	SVM	81.3
		Bakry and Elgammal [2]	SVM	84.84
		Zhou et al. [78]	SVM	85.6
		Pei et al. [51]	SVM	89.7
CUAVE	Hand crafted	DCT [20]	HMM	64
		AAM [48]	HMM	75.7
		Lucey and Sridharan [37]	HMM	77.08
		Visemic AAM [49]	HMM	83
	Deep learning	Ngiam et al. [41]	SVM	66.7
		Srivastava et al. [62]	SVM	69.0

list the performance of these methods for three popular publicly available visual speech corpora in Table 1. The table shows that graph-based and deep learning-based methods generally perform better than handcrafted feature-based approaches. Although some geometric-based handcrafted features [37, 48, 49] achieved more accurate results compared to the graph-based and deep learning-based methods, it is required that the landmarks on the facial area are laboriously labelled beforehand. On this basis, the VSR research community generally recognises that graph-based and deep learning-based methods should be the focus of future research.

Most graph-based and deep learning-based methods have been developed in an attempt to pose lipreading as a classification problem. However, in order to employ VSR for connected and continuous speech applications, the VSR problem should be tackled in a similar way to a speaker-independent acoustic speech recognition task [29]. In terms of continuous speech recognition, instead of extracting holistic visual features from the videos, visual information needs to be represented in a frame-wise manner-that is, the spatio-temporal visual features should be extracted frame by frame, and the temporal dynamic information needs to be captured by the classifiers (e.g., HMM). Given that acoustic modelling for speech recognition using deep learning techniques has been extensively investigated by the speech

community, and given that some of these systems have already been commercialised in recent years [26], it is worth investigating whether these methods can be used for VSR.

Another challenge in the area of VSR is that a large-scale and comprehensive data corpus needs to be available. Although there are a large number of data corpora available for VSR research, all of the existing ones cannot satisfy the ultimate goal, which is to build a practical lipreading system that can be used in real-life applications. That is, in order to treat the VSR problem in a way that is similar to continuous speech recognition, which needs to capture the temporal dynamics of the data (e.g., by using HMMs), a large-scale audio-visual data corpus needs to be established, as this will provide visual speech in the same context as audio speech. Currently, popular benchmark corpora such as AVLetters [38], CUAVE [50] and OuluVS [74] are not fully useful because they are limited in both speaker number and speech content. In addition, some large-scale data corpora such as AVTIMIT [23], IBMSR [37], IBMIH [28] are not publicly accessible. Although the publicly available XM2VTSDB [39] has 200 speakers, the speech is limited to simple sequences of isolated word and digit utterances. A large-scale and comprehensive data corpus called AusTalk was recently created [5, 6, 67, 72]. AusTalk is a large 3D audio-visual database of spoken Australian English recorded at 15 different locations in all states and territories of Australia. The contemporary voices of one thousand Australian English speakers of all ages have been recorded in order to capture the variability in their accent, linguistic characteristics and speech patterns. To satisfy a variety of speech-driven tasks, several types of data have been recorded, including isolated words, digit sequences and sentences. Given that the AusTalk data corpus is a relatively new dataset, only a few works have used this data corpus to date [68–70]. A comprehensive review on the availability of data corpora can also be found in [79].

This chapter reviewed the recent advances in the area of visual speech feature representation. One can conclude from this survey that graph-based and deep learning-based feature representations are generally considered state-of-the-art. Instead of directly using handcrafted visual features for the VSR task, handcrafted visual feature extraction methods are widely used during the pre-processing phase before the extraction of visual features that are finally used for graph-based and deep learning techniques. Despite the exciting recent achievements by the VSR community, several challenges still need to be addressed before a system is developed that can fulfil the specifications of real-life applications. We have summarised the major challenges and proposed possible solutions in this chapter.

References

1. T.R. Almaev, M.F. Valstar, Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition, in *Proceedings of Humaine Association Conference on Affective Computing and Intelligent Interaction* (IEEE, Geneva, 2013), pp. 356–361

2. 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, Washington, 2013), pp. 684–691
3. Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, Greedy layer-wise training of deep networks. Adv. Neural Inf. Process. Syst. **19**, 153 (2007)
4. C. Bregler, Y. Konig, eigenlips for robust speech recognition, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2 (IEEE, Washington, 1994), pp. 669–672
5. 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 *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus* (2009), pp. 96–107
6. 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 Twelfth Annual Conference of the International Speech Communication Association* (2011)
7. H.E. Cetingul, Y. Yemez, E. Erzin, A.M. Tekalp, Discriminative analysis of lip motion features for speaker identification and speech-reading. IEEE Trans. Image Process. **15**(10), 2879–2891 (2006)
8. Y. Cheung, X. Liu, X. You, A local region based approach to lip tracking. Pattern Recogn **45**(9), 3336–3347 (2012)
9. N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1 (IEEE, Washington, 2005), pp. 886–893
10. M. Dantone, J. Gall, G. Fanelli, L. van Gool, Real-time facial feature detection using conditional regression forests, in *Proceedings of International Conference on Computer Vision and Pattern Recognition* (2012)
11. P. Delmas, P. Coulon, V. Fristot, Automatic snakes for robust lip boundaries extraction, in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, vol. 6 (IEEE, Washington, 1999), pp. 3069–3072
12. L. Deng, D. Yu, *Deep Learning: Methods and Applications* (Now Publishers, Boston, 2014)
13. S. Dupont, J. Luettin, Audio-visual speech modeling for continuous speech recognition. IEEE Trans. Multimedia **2**(3), 141–151 (2000)
14. A. Elgammal, C.S. Lee, Separating style and content on a nonlinear manifold, in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, vol. I (IEEE, Washington, 2004), pp. 478–485
15. V. Estellers, M. Gurban, J. Thiran, On dynamic stream weighting for audio-visual speech recognition. IEEE Trans. Audio Speech Lang. Process. **20**(4), 1145–1157 (2012)
16. J. Gehring, Y. Miao, F. Metze, A. Waibel, Extracting deep bottleneck features using stacked auto-encoders, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (IEEE, Washington, 2013), pp. 3377–3381
17. Y. Gizatdinova, V. Surakka Feature-based detection of facial landmarks from neutral and expressive facial images. IEEE Trans. Pattern Anal. Mach. Intell. **28**(1), 135–139 (2006)
18. J.N. Gowdy, A. Subramanya, C. Bartels, J. Bilmes, Dbn based multi-stream models for audio-visual speech recognition, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1 (IEEE, Washington, 2004), pp. 993–996
19. A. Graves, Ar. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in *Proceedings ofIEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, Washington, 2013), pp. 6645–6649
20. M. Gurban, J. Thiran, Information theoretic feature extraction for audio-visual speech recognition. IEEE Trans. Signal Process. **57**(12), 4765–4776 (2009)
21. M. Hayat, M. Bennamoun, S. An, Learning non-linear reconstruction models for image set classification, in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 1915–1922

22. M. Hayat, M. Bennamoun, S. An, Deep reconstruction models for image set classification. IEEE Trans. Pattern Anal. Mach. Intell. **37**(4), 713–727 (2015)
23. T.J. Hazen, K. Saenko, C.H. La, J.R. Glass, A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments, in *Proceedings of the 6th international conference on Multimodal interfaces* (ACM, New York, 2004), pp. 235–242
24. G. Hinton, Training products of experts by minimizing contrastive divergence. Neural Comput. **14**(8), 1771–1800 (2002)
25. G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)
26. G. Hinton, L. Deng, D. Yu, G.E. Dahl, Ar. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath et al., Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process. Mag. **29**(6), 82–97 (2012)
27. J. Huang, B. Kingsbury, Audio-visual deep learning for noise robust speech recognition, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, Washington, 2013), pp. 7596–7599
28. J. Huang, G. Potamianos, J. Connell, C. Neti, Audio-visual speech recognition using an infrared headset. Speech Comm. **44**(1), 83–96 (2004)
29. X. Huang, A. Acero, H.W. Hon et al., *Spoken Language Processing* (Prentice Hall, Englewood Cliffs, 2001)
30. B. Jiang, M.F. Valstar, M. Pantic, Action unit detection using sparse appearance descriptors in space-time video volumes, in *Proceedings of IEEE International Conference on Automatic Face & Gesture Recognition* (IEEE, Washington, 2011), pp. 314–321
31. S. Lankton, A. Tannenbaum, Localizing region-based active contours. IEEE Trans. Image Process. **17**(11), 2029–2039 (2008)
32. Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
33. Y. LeCun, K. Kavukcuoglu, C. Farabet, Convolutional networks and applications in vision, in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems* (IEEE, Washington, 2010), pp. 253–256
34. H. Lee, C. Ekanadham, A.Y. Ng, Sparse deep belief net model for visual area V2, in Proceedings of Adv. Neural Inf. Process. Syst., 873–880 (2008)
35. M. Li, Y. Cheung, Automatic lip localization under face illumination with shadow consideration. Signal Process. **89**(12), 2425–2434 (2009)
36. A. Liew, S. Leung, W. Lau, Lip contour extraction from color images using a deformable model. Pattern Recogn. **35**(12), 2949–2962 (2002)
37. P. Lucey, S. Sridharan, Patch-based representation of visual speech, in *Proceedings of the HCSNet workshop on Use of vision in human-computer interaction*, vol. 56 (Australian Computer Society, Inc., 2006), pp. 79–85
38. I. Matthews, T.F. Cootes, J.A. Bangham, S. Cox, R. Harvey, Extraction of visual features for lipreading. IEEE Trans. Pattern Anal. Mach. Intell. **24**(2), 198–213 (2002)
39. K. Messer, J. Matas, J. Kittler, J. Luettin, G. Maitre, Xm2vtsdb: The extended m2vts database, in *Proceedings of Second International Conference on Audio and Video-based Biometric Person Authentication*, vol. 964, Citeseer (1999), pp. 965–966
40. Ar. Mohamed, G.E. Dahl, G. Hinton, Acoustic modeling using deep belief networks. IEEE Trans. Audio Speech Lang. Process. **20**(1), 14–22 (2012)
41. J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng, Multimodal deep learning, in *Proceedings of the 28th International Conference on Machine Learning* (2011), pp. 689–696
42. Q. Nguyen, M. Milgram, T. Nguyen, Multi features models for robust lip tracking, in *Proceedings of International Conference on Control, Automation, Robotics and Vision* (IEEE, Washington, 2008), pp. 1333–1337
43. K. Noda, Y. Yamaguchi, K. Nakadai, H.G. Okuno, T. Ogata, Audio-visual speech recognition using deep learning. Appl. Intell. **42**, 1–16 (2014)
44. K. Noda, Y. Yamaguchi, K. Nakadai, H.G. Okuno, T. Ogata, Lipreading using convolutional neural network, in *Proceedings of INTERSPEECH* (2014), pp. 1149–1153

45. T. Ojala, M. Pietikäinen, D. Harwood, A comparative study of texture measures with classification based on featured distributions. Pattern Recogn. **29**(1), 51–59 (1996)
46. E. Ong, R. Bowden, Learning sequential patterns for lipreading, in Proceedings of the 22nd British Machine Vision Conference (2011), pp. 55.1–55.10
47. K. Paleček, Extraction of features for lip-reading using autoencoders, in *Speech and Computer* (Springer, Berlin, 2014), pp. 209–216
48. G. Papandreou, A. Katsamanis, V. Pitsikalis, P. Maragos, Multimodal fusion and learning with uncertain features applied to audiovisual speech recognition, in *Proceedings of IEEE 9th Workshop on Multimedia Signal Processing* (IEEE, Washington, 2007), pp. 264–267
49. G. Papandreou, A. Katsamanis, V. Pitsikalis, P. Maragos, Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. IEEE Trans. Audio Speech Lang. Process. **17**(3), 423–435 (2009)
50. E.K. Patterson, S. Gurbuz, Z. Tufekci, J. 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, Washington, 2002), pp. 2017–2020
51. Y. Pei, T.K. Kim, H. Zha, Unsupervised random forest manifold alignment for lipreading, in *Proceedings of IEEE International Conference on Computer Vision* (IEEE, Washington, 2013), pp. 129–136
52. R. Poppe, A survey on vision-based human action recognition. Image Vis. Comput. **28**(6), 976–990 (2010)
53. G. Potamianos, H.P. Graf, E. Cosatto, An image transform approach for hmm based automatic lipreading, in *Proceedings of International Conference on Image Processing* (IEEE, Washington, 1998), pp. 173–177
54. G. Potamianos, C. Neti, G. Gravier, A. Garg, A.W. Senior, Recent advances in the automatic recognition of audiovisual speech. Proc. IEEE **91**(9), 1306–1326 (2003)
55. G. Potamianos, C. Neti, J. Luettin, I. Matthews, Audio-visual automatic speech recognition: an overview. Issues Vis. Audio-Vis. Speech Process. **22**, 23 (2004)
56. M. Ranzato, F.J. Huang, Y.L. Boureau, Y. LeCun, Unsupervised learning of invariant feature hierarchies with applications to object recognition, in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, Washington, 2007), pp. 1–8
57. K. Saenko, K. Livescu, J. Glass, T. Darrell, Multistream articulatory feature-based models for visual speech recognition. IEEE Trans. Pattern Anal. Mach. Intell. **31**(9), 1700–1707 (2009)
58. T.N. Sainath, B. Kingsbury, B. Ramabhadran, Auto-encoder bottleneck features using deep belief networks, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (IEEE, Washington, 2012), pp. 4153–4156
59. R. Salakhutdinov, G.E. Hinton, Deep boltzmann machines, in *Proceedings of International Conference on Artificial Intelligence and Statistics* (2009), pp. 448–455
60. P. Scanlon, R. Reilly, Feature analysis for automatic speechreading, in *Proceeding of IEEE Fourth Workshop on Multimedia Signal Processing* (IEEE, Washington, 2001), pp. 625–630
61. X. Shao, J. Barker, Stream weight estimation for multistream audio–visual speech recognition in a multispeaker environment. Speech Comm. **50**(4), 337–353 (2008)
62. N. Srivastava, R. Salakhutdinov, Multimodal learning with deep boltzmann machines. J. Mach. Learn. Res. **15**, 2949–2980 (2014)
63. N. Srivastava, R.R. Salakhutdinov, Multimodal learning with deep boltzmann machines, in *Advances in neural information processing systems* (2012), pp. 2222–2230
64. D. Stewart, R. Seymour, A. Pass, J. Ming, Robust audio-visual speech recognition under noisy audio-video conditions. IEEE Trans. Cybern. **44**(2), 175–184 (2014)
65. 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, Washington, 2004)
66. J. Su, S. Kurtek, E. Klassen, A. Srivastava et al., Statistical analysis of trajectories on riemannian manifolds: bird migration, hurricane tracking and video surveillance. Ann. Appl. Stat. **8**(1), 530–552 (2014)

67. 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)
68. C. Sui, R. Togneri, S. Haque, M. Bennamoun, Discrimination comparison between audio and visual features, in *Proceedings of the Forty Sixth Asilomar Conference on Signals, Systems and Computers* (IEEE, Washington, 2012), pp. 1609–1612
69. C. Sui, M. Bennamoun, R. Togneri, S. Haque, A lip extraction algorithm using region-based ACM with automatic contour initialization, in *Proceedings of IEEE Workshop on Applications of Computer Vision* (IEEE, Washington, 2013), pp. 275–280
70. C. Sui, R. Togneri, M. Bennamoun, Extracting deep bottleneck features for visual speech recognition, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing* (IEEE, Washington, 2015)
71. P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.A. Manzagol, Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. **11**, 3371–3408 (2010)
72. 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
73. D. Yu, M.L. Seltzer, Improved bottleneck features using pretrained deep neural networks, in *Proceedings of INTERSPEECH* (2011)
74. G. Zhao, M. Barnard, M. Pietikainen, Lipreading with local spatiotemporal descriptors. IEEE Trans. Multimedia **11**(7), 1254–1265 (2009)
75. G. Zhao, X. Huang, Y. Gizatdinova, M. Pietikäinen, Combining dynamic texture and structural features for speaker identification, in *Proceedings of the 2nd ACM workshop on Multimedia in forensics, security and intelligence* (ACM, 2010), pp. 93–98
76. G. Zhao, T. Ahonen, J. Matas, M. Pietikainen, Rotation-invariant image and video description with local binary pattern features. IEEE Trans. Image Process. **21**(4), 1465–1477 (2012)
77. Z. Zhou, G. Zhao, M. Pietikainen, Towards a practical lipreading system, in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, Washington, 2011), pp. 137–144
78. Z. Zhou, X. Hong, G. Zhao, M. Pietikainen, A compact representation of visual speech data using latent variables. IEEE Trans. Pattern Anal. Mach. Intell. **36**(1), 181–187 (2014)
79. Z. Zhou, G. Zhao, X. Hong, M. Pietikäinen, A review of recent advances in visual speech decoding. Image Vis. Comput. **32**(9), 590–605 (2014)