Lip Reading in Profile

Joon Son Chung
http://www.robots.ox.ac.uk/~joon
Andrew Zisserman
http://www.robots.ox.ac.uk/~az

Visual Geometry Group
Department of Engineering Science
University of Oxford
Oxford, UK

Abstract

There has been a quantum leap in the performance of automated lip reading recently due to the application of neural network sequence models trained on a very large corpus of aligned text and face videos. However, this advance has only been demonstrated for frontal or near frontal faces, and so the question remains: can lips be read in profile to the same standard?

The objective of this paper is to answer that question. We make three contributions: first, we obtain a new large aligned training corpus that contains profile faces, and select these using a face pose regressor network; second, we propose a curriculum learning procedure that is able to extend SyncNet [10] (a network to synchronize face movements and speech) progressively from frontal to profile faces; third, we demonstrate lip reading in profile for unseen videos.

The trained model is evaluated on a held out test set, and is also shown to far surpass the state of the art on the OuluVS2 multi-view benchmark.

1 Introduction

Lip reading (or visual speech recognition) is the ability to understand speech using only visual information. As with many perception tasks, machine based lip reading has seen a tremendous increase in performance due to the availability of large scale datasets and the application of neural network based models using deep learning. Lip reading examples include word spotting in continuous speech [9], phrase recognition [3], and sentence level transcribing of continuous speech [8].

However, these recent works have only considered frontal or near-frontal faces, most probably for two reasons: first *availabity*: most video material has mainly near-frontal faces; and second *technological*: until recently, profile face detectors and profile landmark detectors were far inferior to their frontal counterparts. In this paper we extend lip reading to profile faces. We are able to do this, in part, because of the availability of a new generation of ConvNet based object category detectors such as [18, 20].

We then ask the question "Can lips be read in profile to the same standard as those in frontal views?". We might expect the answer to be 'no', since profile views contain less information – the teeth and tongue cannot be seen to the same extent, for example. We investigate this question by generating a new dataset containing copious faces in profile to train and test on, and use this to train a multi-view lip reading network for continuous speech at the sentence level. We also evaluate the network on a recently released public benchmark dataset for multi-view lip reading [2].

Why are profiles important for lip reading? First, a machine that can lip read opens up a host of applications: 'dictating' instructions or messages to a phone in a noisy environment; reading conversations at a distance; or reading archival video without sound. Not having profiles limits this applicability. Second, and quite tantalising, it will become possible to know what HAL lip read in the film '2001: A Space Odyssey' (where the conversation is in profile view).

In detail, we make the following contributions: (i) we obtain a new large aligned corpus, MV-LRS, that contains profile faces selected using a face pose regressor network (Section 2); (ii) we propose a curriculum learning procedure that is able to extend SyncNet [10] progressively from frontal to profile faces (Section 3); and (iii) we train a single sequence-to-sequence model that is able to decode visual sequences across all views, and demonstrate lip reading in profile for unseen videos (Section 5). SyncNet is an essential component of the system: it is used both in building the dataset (for synchronization and for active speaker detection), and it provides the features for the sequence-to-sequence model. Previously [10] it had been applied only on frontal faces, and the extension to profiles here is a necessary, but challenging, step.

The performance of the trained model far exceeds the existing methods on the multi-view test set, and is also shown to surpass the state of the art on the OuluVS2 multi-view benchmark.

1.1 Related works

Research on automatic lip reading has a long history. A large portion of the work has been based on hand-crafted methods, and a comprehensive survey of these methods is given in [28]. We will not review these in detail here.

There have been phenomenal improvements to the performance of lip reading models in recent months, benefitting from advances in deep learning [15, 24], and the ability to obtain and process large scale datasets. These works have shown promising results on transcribing phrases [3] and sentences [8] into words, and have exceeded human performance on their respective datasets.

However, for the most part, existing work has only considered frontal or near-frontal views. The only notable exceptions are the works on the small OuluVS2 multi-view lip reading dataset [2], such as Saitoh *et al.* [22] and Lee *et al.* [16], where the task is to classify visual sequences into one of the 10 phrases in the dataset (*e.g.* 'hello' and 'thank you'). To a large part, this concentration on frontal faces is due to the lack of large-scale datasets that contain profile faces, as is evident in Table 1 which compares existing lip reading datasets.

Name	**Type**	**View**	**Vocab**	**# Utterances**
OuluVS2 [2]	Fixed phrases	0° - 90°	10	3,640
GRID [11]	Phrases	0°	51	33,000
LRW [9]	Words	0° - 30° †	500	500,000
LRS [8]	Sentences	0° - 30° †	17,428	118,116
MV-LRS	Sentences	0° - 90° †	14,960	74,564

Table 1: Comparison of existing datasets. 0° indicates frontal faces, and † indicates that angles are approximate.

2 Dataset collection

We propose a multi-stage strategy to automatically collect a large scale dataset for multi-view lip reading. The dataset is based on the Lip Reading Sentences dataset (LRS) [8], but it contains videos of talking faces covering all views, from frontal to profile.

Whereas the LRS dataset consists of videos taken from mainly broadcast news, we choose a wider range of programs including dramas and factual programs where people engage in conversations with one another, and are therefore more likely to be pictured from the side.

The data preparation pipeline is closely related to [9], and includes the following stages: (i) detect all faces and combine these into face tracks; (ii) temporally align the audio with the subtitles on TV; (iii) correct the audio-to-video synchronisation. This in turn provides the time alignment between the visual face sequence and the words spoken; and, (iv) determine which face is speaking the words (active speaker detection). The first two stages are described in more detail below. Stages (iii) and (iv) employ SyncNet, and these are described in section 3.

2.1 Face tracking

CNN face detector based on the Single Shot MultiBox Detector (SSD) [18] is used to detect face appearances in the individual frames. Unlike the HOG-based detector [14] used by previous works, the SSD detects faces from all angles, and shows a more robust performance whilst being faster to run.

The shot boundaries are determined by comparing color histograms across consecutive frames [17]. Within each shot, face tracks are generated from face detections based on their positions, as featured-based trackers such as KLT [19] often fail when there are extreme changes in viewpoints.

2.2 Channel alignment

The goal is to find the time alignment between the visual face sequence and the words in the subtitle. This is done in two stages: (1) aligning audio to text; (2) aligning video to audio.

Audio to text alignment. TV subtitles are not always in sync with the words being spoken, as they are often typed live. As done in previous works [9], the Penn Forced Aligner [26] is used to align the subtitle to the audio speech; and the force-aligned subtitles are double-checked against a transcript given by a commercial speech recognition software.

Audio to video alignment. On broadcast television, the lip-sync (audio-to-video synchronisation) errors of up to a few hundred milliseconds are common, due to transmission delays, etc. This would result in time offsets between the aligned word and the visual face sequence. The lip-sync error is corrected using SyncNet, described in section 3.

2.3 Facial pose estimation

In order to faciliate the testing of the multi-view model, we divide the data into five pose categories based on the yaw-rotation of the face: (1) left profile; (2) left three-quarter; (3) frontal; (4) right three-quarter; (5) right profile. This is done using a ResNet-based pose regressor, trained on the CASIA-WebFace dataset [25]. The network has been trained to classify cropped face images into one of the above five categories. Examples belonging to each class are given in Figure 1.

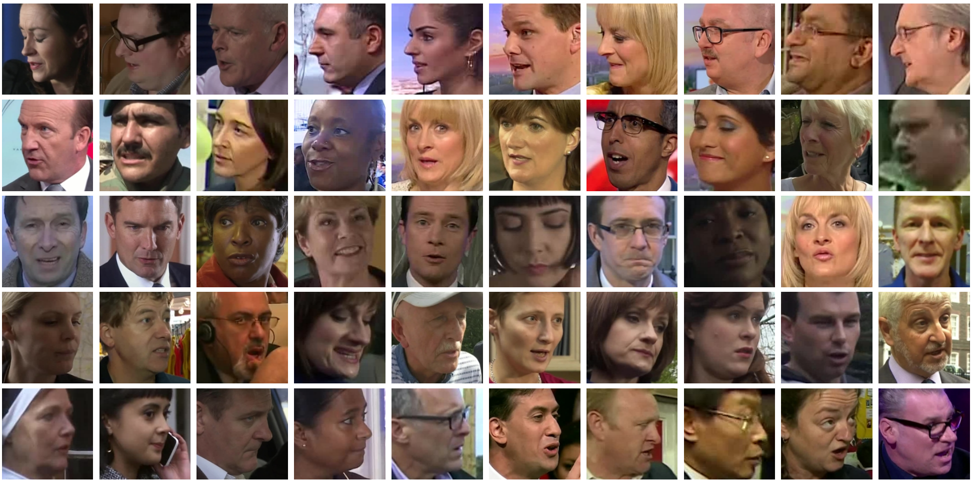

Figure 1: Face detections examples from the MV-LRS dataset. **Top row**: left profile; **2nd row**: left three-quarter; **3rd row**: frontal; **4th row**: right three-quarter; **Bottom row**: right profile.

2.4 Data statistics

Set	Dates	# Sentences	Vocab
Train	01/2010 - 12/2015	67,793	14,440
Val	01/2016 - 02/2016	2,352	4,330
Test	03/2016 - 09/2016	4,429	4,375
All		74,574	14,960

Table 2: **The Multi-View Lip Reading Sentences (MV-LRS) dataset.** Division of training, validation and test data; and the number of utterances and vocabulary size of each partition.

The videos are divided into train, validation and test sets according to date, and in particular, the dates used for the split are the same as the LRS dataset [8]. This is so that the users of the dataset can co-train on the larger LRS dataset, as some of the videos may overlap.

3 Multi-view SyncNet

The SyncNet architecture proposed in [10] is used for three purposes in this paper: first, to synchronize the audio and lip motion in the video sequence; second, for active speaker detection; and third to generate the features for the sequence-to-sequence model. The first and second are used in the dataset construction of Section 2.

In this section, we first review the SyncNet of [10], and then extend SyncNet from the (originally) frontal to the profile faces required for this paper, using a curriculum learning strategy described in Section 3.2.

3.1 SyncNet review

SyncNet learns a joint embedding between the sound and the mouth motions from unlabelled data. The network consists of two asymmetric streams for audio and video, which are described below.

Audio representation. The input audio data is MFCC values. The features are computed at a sampling rate of 100Hz, giving 20 time steps for a 0.2-second input signal.

Video representation. The original SyncNet ingests five precisely aligned lip images, given that the facial landmarks are clearly visible from the front; however, the landmarks are not well-defined in the multi-view case. Here, the multi-view SyncNet takes a larger image region (the whole face bounding box), and hence a larger image resolution of 224×224.

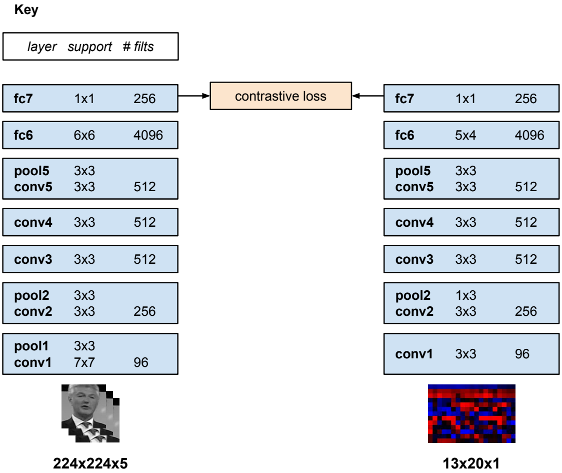

Figure 2: Multi-view SyncNet architecture.

Architecture. Both streams in SyncNet are based on the standard VGG-M [6] architecture. The modified network shares the underlying layer structure of the original SyncNet, but the visual stream has slightly different filter sizes to accommodate the larger input size. The layer configurations are shown in Figure 2.

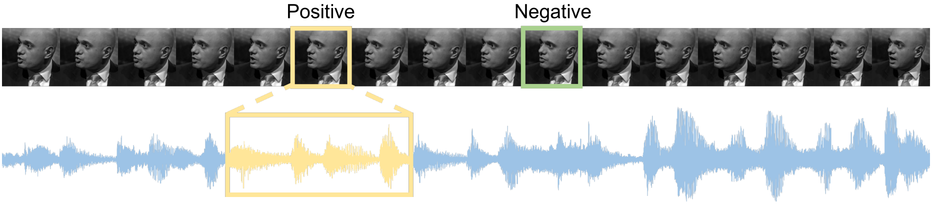

Figure 3: Sampling strategy for training SyncNet.

Training protocol. We use a curriculum learning strategy described in Section 3.2; otherwise the training protocol follows that of [10] – positive audio-video pairs are taken from corresponding frames in validated facetracks, and negative audio-video pairs are generated by randomly selecting non-corresponding frames from the same face track. The sampling strategy is shown in Figure 3. The two-stream network is trained with a contrastive loss to minimise the distance between features for positive pairs, and maximise the distance for negatives.

3.2 Curriculum learning

The training of the original SyncNet used the assumption that the majority of faces in the dataset are speaking. Whilst this may be the case for the news programmes it was trained on (Figure 4 left), this assumption cannot be used to bootstrap the multi-view model as there would be too much noise (in the form of non-speaking faces) for the network to learn relevant information. For example in a scene such as Figure 4 right, only one of these faces would be speaking at any one point. To circumvent this problem, we start with the frontal SyncNet trained on the news programmes, and adopt a curriculum learning approach that gradually increases the working angle of the active speaker detection system.

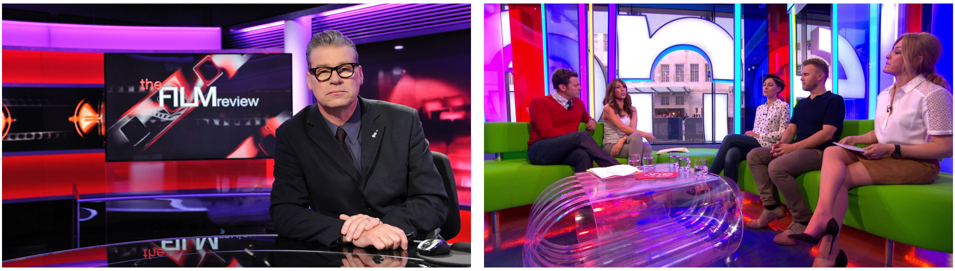

Figure 4: **Left:** Still image from ‘BBC News’; **Right:** Still image from ‘The One Show’.

Stage 1. Frontal view. The first stage is to determine speaking and non-speaking face sequences for the frontal faces (view 3 from Section 2.3). Facial landmarks are determined using the regression-tree based method of [13]. The landmarks are used to align and crop the lip region; active speaker detection is performed on all tracks using the frontal-only SyncNet on the aligned lip images. The new network is trained on the active speaker images using the full face image (instead of the aligned lips).
Stage 2. Three-quarter view. The network trained in Stage 1 is used to determine the active speaker on the three-quarter view (views 2 and 4) face tracks. The speaking tracks from these views are added to the training data; and the synchronisation network is re-trained.
Stage 3. Profile view. As before, the network in Stage 2 is used to perform speaker detection on the profile view (views 1 and 5) tracks. The speaking tracks are added to the training data and the network is re-trained.
Evaluation. We report Equal Error Rates on the labelled validation set in Table 3. The data is in the same format as used in training – the correct audio-video pairs for positives, and artificially shifted audio for negatives. Note that not every 0.2-second sample contains discriminative information even within a labelled segment of speech (*e.g.* the person might be taking a breath), but it nonetheless illustrates the performance gained from the curriculum training.
Discussion. Using this method, we are able to train a two-stream network that learns an embedding of the audio and the lip motion, and provides a robust method of correcting the lip-sync error, and determining the active speaker in multi-speaker scenes.

The method does not require any annotation of the training data and allows almost any web video to be used as training data, so the cost of obtaining the training data is minimal.

	MV SyncNet	SyncNet
Frontal	13.6%	**13.2%**
Three-quarter	**14.8%**	17.1%
Profile	**16.2%**	21.7%

Table 3: **Equal Error Rates** on the validation set, using single 0.2-second samples. Lower is better.

As shown in [10], the visual stream of this network generates excellent features for the task of lip reading – on the LRW [9] and OuluVS2 [2] datasets, single-layer classifiers trained on the SyncNet features have outperformed networks trained end-to-end on the task. This is presumably because the SyncNet is trained on a near-infinite amount of audio-visual data, whereas this is not feasible for lip reading.

4 MV-WAS Architecture

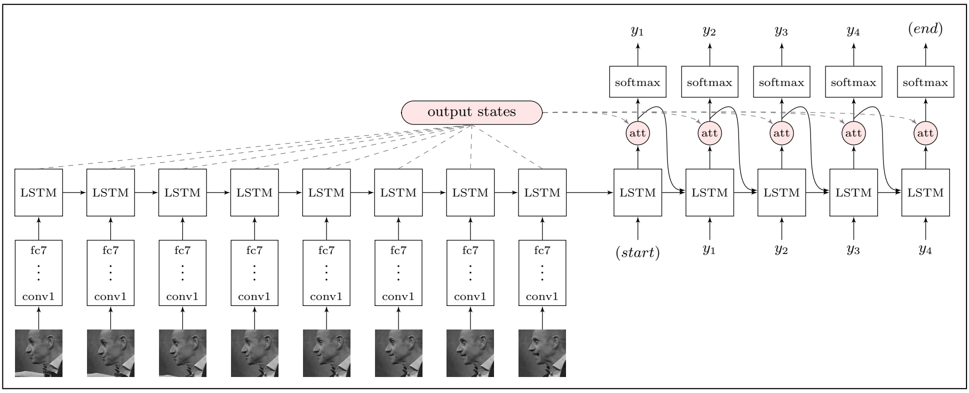

Figure 5: *MV-WAS* architecture.

The Multi-view Watch, Attend and Spell (MV-WAS) model is based on the WAS model of [8], without the second attention mechanism and the audio encoder. The network configuration is shown in Figure 5. The model consists of two key modules: the image encoder and the character decoder, described in the following paragraphs:

Image encoder. The image encoder consists of the convolutional part that generates image features for every input timestep, and the recurrent part that produces the fixed-dimensional state vector and a set of output vectors.

The convolutional layer configurations are based on the VGG-M model [6], as it is memory-efficient and fast to train compared to deeper model such as VGG-16 [23] and ResNet [12], whilst still showing good classification performance on ImageNet [21].

To prevent overfitting and for computational efficiency, the convolutional layer weights (*conv1* to *conv5*) are fixed to that of the multi-view SyncNet. Memory-efficiency is important here as a large number of images (# timesteps × batch size) must be passed through the ConvNet at every iteration, and in particular the input images are significantly larger than that used by the original WAS network [8].

The encoder LSTM network ingests the output features produced by the ConvNet at every timestep, and generates a fixed-dimensional state vector at the end of the sequence, and an output vector at every timestep, to be read by the attention decoder.

Character decoder. The decoder module uses a LSTM transducer [4, 5, 7] to produce a probability distribution over the next character conditioned on the inputs and the previous characters, one character at a time. This transducer is based on the implementation of [5] and will not be repeated here in detail.

Implementation details. Our implementation is based on the TensorFlow library [1] and trained on a NVIDIA GeForce GTX 1080 GPU. The network is trained with dropout, and a batch size of 64 was used.

5 Experiments

5.1 Evaluation on MV-LRS

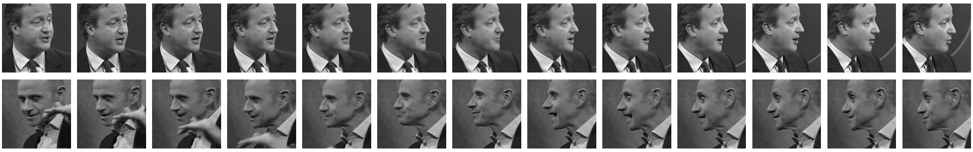

Figure 6: Example video frames from sentences in the MV-LRS dataset.

Training. The MV-WAS model is trained using the curriculum learning approach described in [8], where the model starts to learn from easier, single-word examples and gradually move to longer sentences. This results in faster training and less overfitting. A single multi-view model is trained (as opposed to separate models for every viewpoint) given that the viewpoint may change within a sentence (as shown in Figure 6), and the amount of data for each viewpoint would be insufficient for training in any case.

We compare performance to the WAS model [8]. This model is pre-trained on the LRS dataset, and we fine-tune the LSTM layers on the multi-view dataset until the validation error stops improving. This is done so that the language model (implicitly learnt in the decoder) adapts to the new corpus that consists of videos from previously unseen genres (*e.g.* dramas).

Evaluation protocol. The performance measures used are consistent with that used in related works [3, 8] – we report the Character Error Rate (CER), the Word Error Rate (WER) and the unigram BLEU measure.

Decoding. The decoding is performed with a beam size of 4.

Viewpoint	**MV-WAS**			**WAS [8]**		
	CER	WER	BLEU†	CER	WER	BLEU†
Frontal	46.5%	56.4%	49.3	**45.5%**	**56.1%**	**50.4%**
Three-quarter	**50.4%**	**59.2%**	**46.1**	55.4%	65.2%	42.5
Profile	**54.4%**	**62.8%**	**42.5**	74.2%	82.6%	26.6

Table 4: Results on the MV-LRS dataset. Lower is better for CER and WER; higher is better for BLEU. †Unigram BLEU with brevity penalty.

Results. Performance measures for all viewpoints are given in Table 4. The profile performance of the **MV-WAS** model far exceeds the frontal-only WAS model fine-tuned on our dataset, and also shows a significant improvement for three-quarter faces. The performance of our model on frontal videos is comparable to that of the frontal-only WAS model. Table 5 gives examples of successfully read sentences.

AND IF YOU LOOK AROUND THE WORLD NOW
BUT BEHIND THE SCENES THERE IS ANOTHER
DESPITE THIS STRONG GESTURE OF PEACE
TENS OF MILLIONS OF CHILDREN ARE LEFT BEHIND
OUR RELATIONSHIP WITH THE REST OF THE WORLD

Table 5: Examples of unseen sentences in profile-view that MV-WAS correctly predicts. The examples are best seen in video format. Please see the online examples.

5.2 Evaluation on OuluVS2

We evaluate the MV-WAS model on the OuluVS2 dataset [2]. The dataset consists of 52 subjects uttering 10 phrases (*e.g*. 'thank you', 'hello', etc.), and has been widely as a benchmark. Here, we assess on a speaker-independent experiment, where 12 specified subjects are reserved for testing.

Training. We use the sequence-to-sequence model pre-trained on the MV-LRS dataset, and fine-tune the LSTM layers on the training portion of the OuluVS2 data. Unlike previous works [16, 22, 27] that use separate models trained for each viewpoint, we only train a single model to classify the phrases at all angles.

Decoding. The decoding is performed with a beam size of 1.

Results. As can be seen in Table 6 our method achieves a strong performance, and sets the new state-of-the-art for the multi-view task.

Method	**Frontal**	**30°**	**45°**	**60°**	**Profile**
Zhou *et al.* [27]	73.0%	75.0%	76.0%	75.0%	70.0%
Lee *et al.* [16]	81.1%	80.0%	76.9%	69.2%	82.2%
Saitoh *et al.* [22]	85.6%	79.7%	80.8%	83.3%	80.3%
MV-WAS (ours)	**91.1%**	**90.8%**	**90.0%**	**90.0%**	**88.9%**

Table 6: **Classification accuracy** on OuluVS2 Short Phrases. Higher is better.

6 Conclusion

We can give a qualified answer to the question posed in the introduction: "Yes, it is possible to read lips in profile, but the standard is inferior to reading frontal faces".

We plan now to increase the size of the dataset further to see if the availablity of more training data will be of benefit. Also, it will be interesting to investigate how deep learning has learnt to select relevant information for each view, and whether different architectures, *e.g*. increasing capacity, will improve performance.

Acknowledgements. Funding for this research is provided by the EPSRC Programme Grant Seebibyte EP/M013774/1.

References

[1] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467 (2016)

[2] Anina, I., Zhou, Z., Zhao, G., Pietikäinen, M.: Ouluvs2: a multi-view audiovisual database for non-rigid mouth motion analysis. In: Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on. vol. 1, pp. 1–5. IEEE (2015)

[3] Assael, Y.M., Shillingford, B., Whiteson, S., de Freitas, N.: Lipnet: Sentence-level lipreading. Under submission to ICLR 2017, arXiv:1611.01599 (2016)

[4] Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. Proc. ICLR (2015)

[5] Chan, W., Jaitly, N., Le, Q.V., Vinyals, O.: Listen, attend and spell. arXiv preprint arXiv:1508.01211 (2015)

[6] Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: Delving deep into convolutional nets. In: Proc. BMVC. (2014)

[7] Chorowski, J.K., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y.: Attention-based models for speech recognition. In: Advances in Neural Information Processing Systems. pp. 577–585 (2015)

[8] Chung, J.S., Senior, A., Vinyals, O., Zisserman, A.: Lip reading sentences in the wild. In: Proc. CVPR (2017)

[9] Chung, J.S., Zisserman, A.: Lip reading in the wild. In: Proc. ACCV (2016)

[10] Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: Workshop on Multi-view Lip-reading, ACCV (2016)

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

[12] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)

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

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

[15] Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: NIPS. pp. 1106–1114 (2012)

[16] Lee, D., Lee, J., Kim, K.E.: Multi-view automatic lip-reading using neural network. In: ACCV 2016 Workshop on Multi-view Lip-reading Challenges. Asian Conference on Computer Vision (ACCV) (2016)

[17] Lienhart, R.: Reliable transition detection in videos: A survey and practitioner's guide. International Journal of Image and Graphics (Aug 2001)

[18] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: Proc. ECCV. pp. 21–37. Springer (2016)

[19] Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proc. of the 7th International Joint Conference on Artificial Intelligence. pp. 674–679 (1981), citeseer.nj.nec.com/lucas81optical.html

[20] Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: NIPS (2016)

[21] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, S., Karpathy, A., Khosla, A., Bernstein, M., Berg, A., Li, F.: Imagenet large scale visual recognition challenge. IJCV (2015)

[22] Saitoh, T., Zhou, Z., Zhao, G., Pietikäinen, M.: Concatenated frame image based cnn for visual speech recognition. In: Asian Conference on Computer Vision. pp. 277–289. Springer (2016)

[23] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)

[24] Sutskever, I., Vinyals, O., Le, Q.: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. pp. 3104–3112 (2014)

[25] Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)

[26] Yuan, J., Liberman, M.: Speaker identification on the scotus corpus. Journal of the Acoustical Society of America 123(5), 3878 (2008)

[27] Zhou, Z., Hong, X., Zhao, G., Pietikäinen, M.: A compact representation of visual speech data using latent variables. IEEE PAMI 36(1), 1–1 (2014)

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