FEATURE EXTRACTION USING MULTIMODAL CONVOLUTIONAL NEURAL NETWORKS FOR VISUAL SPEECH RECOGNITION

Eric Tatulli, Thomas Hueber

CNRS / Univ. Grenoble-Alpes / GIPSA-lab, Grenoble, France

ABSTRACT

This article addresses the problem of continuous speech recognition from visual information only, without exploiting any audio signal. Our approach combines a video camera and an ultrasound imaging system for monitoring simultaneously the speaker's lips and the movement of the tongue. We investigate the use of convolutional neural networks (CNN) to extract visual features directly from the raw ultrasound and video images. We propose different architectures among which a multimodal CNN processing jointly the two visual modalities. Combined with an HMM-GMM decoder, the CNN-based approach outperforms our previous baseline based on Principal Component Analysis. Importantly, the recognition accuracy is only 4% lower than the one obtained when decoding the audio signal, which makes it a good candidate for a practical visual speech recognition system.

Index Terms— Visual Speech Recognition, Convolutional Neural Networks, Deep Learning.

1. INTRODUCTION

The use of visual information in voice-based human-computer interface (HCI) has been investigated in many studies. In this article, we focus on *visual-only speech recognition* (VSR), that is speech recognition without exploiting any audio signal. Since the lips movements provide only a partial information on speech articulation, we have investigated in [1] the use of medical ultrasound imaging to capture also the movement of the tongue, with a probe placed beneath the speaker's chin, as shown in Fig. 1. Such a system can be referred to as a *silent speech interface* [2], since it should enable oral speech communication without the necessity to emit any audible sound. It will allow the design of confidential, non-disturbing, and noise-robust voice-based HCI.

Most studies on VSR (mainly from lips gestures only), decompose this problem into two consecutive stages: the extraction of visual features from raw images, and the classification itself. For the feature extraction stage, many techniques have been proposed among which active shape model [3], active appearance model [4], discrete cosine transform [5], and principal components analysis (PCA) [6]. Similarly to audio-based speech recognition (ASR), the classification stage is often addressed using models able to take explicitly into account the speech dynamics. Among other approaches, graphical models such as hidden Markov model (HMM) [7], coupled-HMM [8], and dynamic Bayesian network [9] have been widely investigated. In our previous work ([1, 10]), we have followed the same pipeline, by using (1) a PCA-based decomposition for encoding both the video and the ultrasound images, and (2) a 2-stream HMM-GMM classifier.

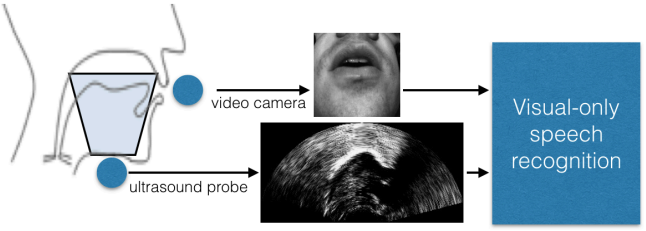

Fig. 1. Overview of the visual speech recognition system from video and ultrasound images.

Deep neural networks have shown in many domains their ability to learn representations directly from the raw data and can be used to extract a set of discriminative features. In the context of image processing, one powerful deep architecture is the so-called *Convolutional Neural Network* (CNN) [11]. The use of CNN for gestures recognition in video has been proposed in few recent studies such as [12, 13, 14]. To the best of our knowledge, the use of CNN in VSR has been investigated in one study [15] for encoding lips images in an isolated word recognition task.

In this article, we investigate the use of CNN to extract visual features directly from the raw ultrasound and video images (Sec. 2). We describe several architectures (Sec. 2.2) among which a multimodal CNN processing jointly the two visual modalities (Sec. 2.3). The proposed method is compared to our baseline based on Principal Component Analysis [10] and to a *golden standard* given by a conventional audio-based speech recognition system (ASR) trained on the same database (Sec. 3). Importantly, we focus in this study on *continuous* speech (as opposed to isolated word) recognition. Results are presented and discussed in Sec. 4.

This work has been supported by the LabExPERSYVAL-Lab (ANR-11-LABX-0025-01) and the European Research Council under the European Communitys Seventh Framework Programme (FP7/2007-2013 Grant Agreement no. 339152, Speech Unit(e)s.

2. VISUAL FEATURE EXTRACTION USING CNN

2.1. Convolutional neural networks

A CNN is a multilayer stack of learning modules well-suited for treating bi-dimensional dataset (i.e. images). CNNs are subclass of neural networks that combine the nonlinear processing of hidden layer neurons with essential properties of weight sharing (over customizable sub-images so-called convolutional filters), pooling and down-sampling. As a consequence, such networks are expected to learn representation of data with increasing levels of abstraction regrouped by semantic similarities. The canonical structure of a CNN [11] contains: 1) a given number of *convolutional layers*, each being divided in four sub-tasks: convolutional filtering, nonlinearity, pooling and sub-sampling, 2) a set of *fully connected layers* with properties identical to that of classical neural networks, 3) a *softmax layer* performing *softmax* function which outputs posterior probabilities for each class.

2.2. Independent processing of the visual modalities

Our first implementation is based on two CNNs, each one processing independently one visual modality (i.e. ultrasound and video). At training stage, the classical gradient-descent back-propagation technique is used to estimate the parameters in a supervised manner, the phonetic labels being used as targets. For each modality, a vector of visual features is extracted from the network by taking the output of the last fully-connected layer (just before the final softmax layer).

2.3. Multimodal architecture

We propose a multimodal CNN processing jointly pairs of video and ultrasound images. This architecure is illustrated in Fig. 2 and consists in the fusion of two canonical CNNs [16]. Importantly, it includes a *fusion layer* combining the ultrasound and video modalities. Such an architecture aims at extracting high-level features from the simulatenous observation of tongue, lips and jaw. As in Section 2.2, the multimodal CNN is trained in a supervised manner, the phonetic labels being used as targets.

Based on this multimodal architecture, we investigated two ways of extracting visual features: 1) extracting one single feature vector at the output of the fusion layer (implementation S3, see Fig. 2, top), and 2) extracting two feature vectors, one for each modality, at the output of the last fully connected layer, before the fusion layer (implementation S4, see Fig. 2, top). Let us mention that the resulting features may be different from the one obtained when considering two CNNs trained separately. Indeed, the two modalities are here tied together. Their parameters are jointly estimated and thus can be mutually influenced.

2.4. HMM-based visual speech recognition

In this study, CNNs are used as feature extractors and are combined with a conventional HMM-GMM phonetic decoder. This architecture allows the introduction of prior linguistic knowledge during the decoding via the use of a pronunciation dictionary and a language model. Such prior knowledge remains of particular interest in the context of visual speech recognition for regularization purposes. As a matter of fact, several sources of information such as the voicing are missing when considering only visual data.

Two strategies can be investigated for combining the ultrasound and video modalities within the HMM-GMM decoder: (1) an *early fusion* strategy in which the feature vectors related to each modality are concatenated together and modeled using a 1-stream HMM-GMM decoder, and (2), a middle fusion strategy based on a 2-stream HMM-GMM decoder where the modalities are combined at the HMM state level. The combination of the two CNN-based feature extraction techniques (independent vs. joint modeling) with these two strategies (early vs. middle fusion) results in 4 VSR architectures. Those architectures are referred to as S1, S2, S3, and S4, and are illustrated in Fig. 2 (bottom).

3. EXPERIMENTS

3.1. Database

Experiments were conducted on the same database used in [10] which contains 488 sentences pronounced by a male French speaker. Ultrasound images (320x240 grayscale images, 60fps) were acquired using the *Terason T3000* medical ultrasound system, with a 128 elements microconvex transducer (3-5 MHz frequency, 140° angle, 7cm penetration depth). Video images of the speaker's face (640x480 grayscale images, 60 fps) were recorded using an industrial CMOS camera. Ultrasound and video sensors were kept fixed with respect to the speaker's head using a stabilization helmet. Visual and audio data were recorded simultaneously using the *Ultraspeech* software [17], in a sound-proof room, and under stable conditions of lightning. French language was described using a set of 34 phonemes. The phonetic transcription of each recorded sentences was extracted automatically and manually post-checked. The temporal boundaries of each phoneme were extracted from the audio signal using a conventional ASR system and a forced-alignment procedure. The phonetic segmentation of the audio signal was then used to label the visual data (since audio, ultrasound and video data are recorded synchronously) and to train the CNNs and HMM-GMM decoders.

3.2. Implementation details

For S1 and S2 systems (independent processing of the two modalities), it appeared that the simplest CNN structure: one

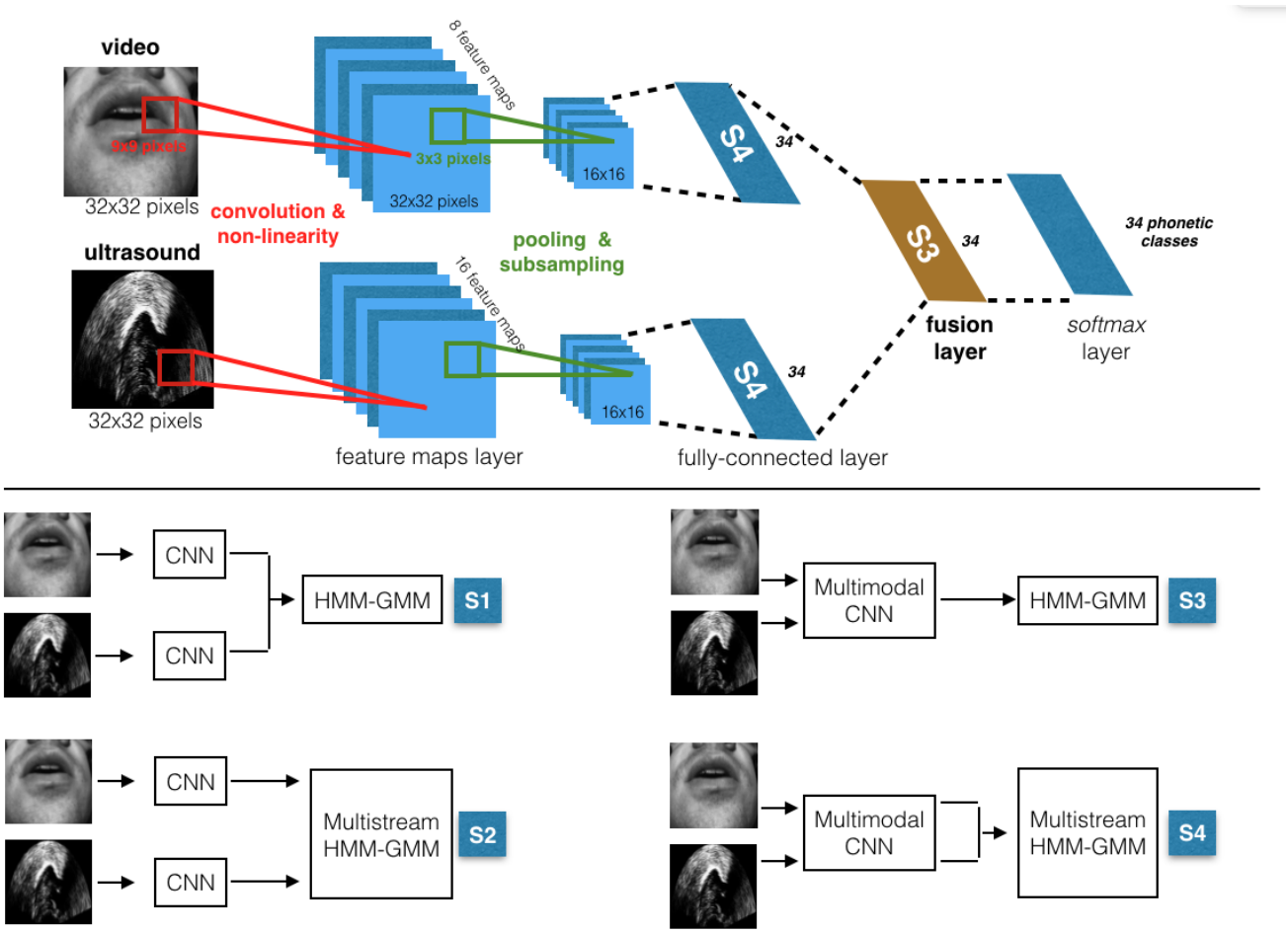

Fig. 2. Top: Architecture of the proposed multimodal convolutional neural network (CNN). Bottom: Schematic representation of the 4 proposed VSR systems (S1,S2,S3, and S4) combining CNN-based feature extraction and HMM-GMM decoding.

convolution layer, one full layer, and one softmax layer with only a moderate number of filters (respectively 16 and 8), provides satisfying results (as discussed in Sec. 4). For S3 and S4 systems (multimodal CNN), the best results were obtained also using a simple structure: one convolution layer, one full layer, one fusion layer and one softmax layer. Given the number of free parameters at play, we do not claim that the proposed architecture is optimal and tuning is likely to improve its performance. For all CNNs, the Rectified Linear Unit (ReLU) non-linearity was used for all convolutional, fusion and full layers. All CNNs were implemented using the *MatconvNet* toolkit [18] and were trained using GPU-acceleration.

All HMM-GMM decoders were built using the HTK toolkit [19] with a standard HMM topology (3 states) and a standard training procedure (tied states and context-dependent triphone modeling). For all experiments, the visual features were modeled together with their first derivatives. At decoding stage, the most likely sequence of phonemes was estimated by decoding the HMM-GMM state posterior probabilities using the Viterbi algorithm (the model insertion penalty was optimized on the training set). For the 2-stream HMM architecture, the weighting parameters used to combine the stream likelihoods were also optimized on the training set. Optimal values were found to be 0.7 for ultrasound and 0.3 for video (a similar result was found in [10]).

Since the present study aim only at probing the ability of the multimodal CNN to process the visual data, recognition experiments were conducted without exploiting prior linguistic information. The performance was measured by calculating the phoneme recognition accuracy T_p defined as $T_p = (N_p - D - S - I)/N_p$ where N_p is the number of phonemes in the test corpus, and D, S and I are respectively the number of deletions, substitutions and insertions. The 95% confidence interval ($\Delta_{95\%}$) of the phonetic recognition rate was computed following [20]. A 8-fold cross-validation was used to refine our statistics, by splitting our corpus into eight subsets, keeping seven subsets for the training and the remaining one for testing (taking into account all the possible permutations).

Table 1. Accuracy of the 4 systems of visual speech recognition based on CNN (S1,S2,S3,S4), and PCA (B1, B2). Comparison with an ASR system trained on the audio stream. For all experiments, the 95%-confidence interval is $\sim 1.5\%$.

	ASR - MFCC					
T_p (%)	84					
	VSR - PCA		VSR - CNN			
	B1	B2	S1	S2	S3	S4
T_p (%)	74.7	73.1	77.8	79.9	75.8	**80.4**

3.3. Baseline

The CNN-based feature extraction approach was compared to a PCA-based approach, as used in our previous studies ([1, 10]). This technique is a slight adaptation of the EigenFaces technique [21] and aims at finding a decomposition basis that best explains the variation of pixel intensity in a set of training frames. At feature extraction stage, resized and normalized video/ultrasound frame is projected onto this basis and the visual features are defined as the D first coordinates, for each stream. The number of coordinates is a free parameter. In our implementation, it is optimized on the training set by keeping the eigenvectors that carry 80% of the variance, which led in our case to $D = 30$ for both video and ultrasound images. Based on this approach, we derive two baseline systems, B1 and B2, where PCA-based features are decoded using either an early fusion strategy (i.e. 1-stream HMM-GMM, as in S1 and S3) or a middle fusion strategy was used (i.e. 2-stream HMM-GMM, as in S2 and S4).

For each experiment, the performance was also compared with the one obtained when considering the audio data (which was recorded simultaneously with the visual data). Considering that audio provides thorough information, we assume that the ASR accuracy gives the upper bound reachable by a VSR system. Audio signal was parametrized using MFCC decomposition (resulting in a vector of 13 static coefficients with their first derivatives, extracted every 5ms). The HMM-GMM decoder was trained using the same procedure as for the VSR systems (3-states, tied-state, context-dependent, triphone models).

4. RESULTS AND DISCUSSION

Results are presented in Table 1. First, CNN-based approaches systematically outperform PCA-based baselines, regardless the strategy used to combine the modalities (early or middle fusion). This demonstrates the potential of the CNNs to extract relevant features from the raw video and ultrasound images. Second, the differences observed between the 4 CNN-based VSR systems are more difficult to interpret. Nonetheless, the following conclusions can be drawn:

1. The middle fusion strategy always outperforms the early fusion since S2>S1 and S4>S3 (surprisingly, it was not the case for the baseline since B1>B2).

2. The best performance was obtained with the multimodal CNN architecture and the middle fusion strategy (S4), with 80.4% accuracy. However, the difference with the system S2 lies within the confidence interval. Therefore, the benefit of considering jointly the two modalities at the feature extraction stage need to be confirmed with additional experiments.

3. The lowest performance was obtained with the system S3 in which the features are extracted at the output of the fusion layer. Among other possible explanations, we can conjecture that the fusion layer operates as a bottle-neck in the network and directly control the dimension of the extracted feature vector. In this study, we empirically set this parameters to 34 in order to match the number of target phonetic classes and limit the number of fully-connected layers before the final softmax layer. Considering the relatively low performance, we infer that the reduction of the dimension (by a factor of 2 in comparison with S1,S2 and S4) is too severe. Hence, the optimization of the size of the fusion layer seems to be a key issue and should be addressed carefully in a future study.

4. With 80.4% accuracy, the best VSR system S4 approaches the ultimate accuracy of 84% derived from audio data. Such performance is very encouraging and makes the multimodal CNN a good candidate for a practical VSR system.

5. CONCLUSIONS

In this article, we investigated the use of CNN for extracting visual features from ultrasound and video images of the tongue and lips. We proposed a multimodal architecture in which the two visual modalities are jointly processed. We derived different systems in which the CNN is used as a feature extractor and is combined with a HMM-GMM decoder. Experiments were conducted on a continuous speech VSR task. Results have demonstrated the potential of the CNN over a previously published baseline. However, further experiments should be conducted to valid the potential benefit of the multimodal architecture over the use of two distincts CNN. Such experiments will be conducted in future studies on a multi-speaker database. Future work will also focus on the design of an end-to-end VSR system (in line with recent work on ASR [22]), combining convolutional layers for processing the raw visual data with a recurrent architecture (as in [23]) to model the dynamics of speech articulation.

6. REFERENCES

[1] Thomas Hueber, Elie-Laurent Benaroya, Gérard Chollet, Bruce Denby, Gérard Dreyfus, and Maureen Stone, "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips," *Speech Communication*, vol. 52, no. 4, pp. 288–300, 2010.

[2] Bruce Denby, Tanja Schultz, Kiyoshi Honda, Thomas Hueber, Jim M Gilbert, and Jonathan S Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[3] Juergen Luettin, Neil A Thacker, and Steve W Beet, "Visual speech recognition using active shape models and hidden markov models," in *Proc. IEEE-ICASSP*, 1996, vol. 2, pp. 817–820.

[4] Astik Biswas, PK Sahu, and Mahesh Chandra, "Multiple cameras audio visual speech recognition using active appearance model visual features in car environment," *International Journal of Speech Technology*, vol. 19, no. 1, pp. 159–171, 2016.

[5] Martin Heckmann, Kristian Kroschel, Christophe Savariaux, Frédéric Berthommier, et al., "DCT-based video features for audio-visual speech recognition.," in *Proc. Interspeech*, 2002, vol. 3, pp. 1925–1928.

[6] C. Bregler and Y. Konig, "Eigenlips for robust speech recognition," in *Proc. IEEE-ICASSP*, 1994, vol. 2, pp. 669–672.

[7] Richard Bowden, Stephen Cox, Richard Harvey, Yuxuan Lan, Eng-Jon Ong, Gari Owen, and Barry-John Theobald, "Recent developments in automated lip-reading," in *Proc. SPIE*, 2013, pp. 89010J–89010J–13.

[8] G. Gravier, G. Potamianos, and C. Neti, "Asynchrony modeling for audio-visual speech recognition," in *Proc. Int. Conf. on Human Lang. Tech. Research*, 2002, pp. 1–6.

[9] John N Gowdy, Amarnag Subramanya, Chris Bartels, and Jeff Bilmes, "DBN based multi-stream models for audio-visual speech recognition," in *Proc. IEEE-ICASSP*, 2004, vol. 1, pp. 993–996.

[10] Thomas Hueber and Gérard Bailly, "Statistical conversion of silent articulation into audible speech using full-covariance HMM," *Computer Speech & Language*, vol. 36, pp. 274–293, 2016.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[12] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt, "Spatio-temporal convolutional sparse auto-encoder for sequence classification.," in *Proc. BMVC*, 2012, pp. 1–12.

[13] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE-CVPR*, 2014, pp. 1725–1732.

[14] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. NIPS*, 2014, pp. 568–576.

[15] Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G Okuno, and Tetsuya Ogata, "Lipreading using convolutional neural network.," in *Proc. Interspeech*, 2014, pp. 1149–1153.

[16] Y. Le Cun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," in *Proc. NIPS*, 1990, pp. 396–404.

[17] Thomas Hueber, Gérard Chollet, Bruce Denby, and Maureen Stone, "Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application," in *Proc. ISSP*, 2008, pp. 365–369.

[18] A. Vedaldi and K. Lenc, "Matconvnet – convolutional neural networks for MATLAB," in *Proc. ACM-Multimedia*, 2015, pp. 689–692.

[19] Steve J Young, "The HTK HMM toolkit: Design and philosophy," Tech. Rep. CUED/F-INFENG/TR 152, Cambridge Univ. Dept. of Eng., 1993.

[20] Gérard Chollet and Claude Montacie, "Evaluating speech recognizers and databases," in *Recent advances in speech understanding and dialog systems*, Heinrich Niemann, M. Lang, and G. Sagerer, Eds., pp. 345–348. Springer, 1988.

[21] Matthew Turk and Alex Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[22] Alex Graves and Navdeep Jaitly, "Towards end-to-end speech recognition with recurrent neural networks.," in *Proc. ICML*, 2014, vol. 14, pp. 1764–1772.

[23] Ming Liang and Xiaolin Hu, "Recurrent convolutional neural network for object recognition," in *Proc. CVPR*, 2015, pp. 3367–3375.