PREDICTING TONGUE MOTION IN UNLABELED ULTRASOUND VIDEO USING 3D CONVOLUTIONAL NEURAL NETWORKS

Chengrui Wu[1], Shicheng Chen[1], Guorui Sheng[1], Pierre Roussel[2], Bruce Denby[1]

[1]Tianjin University, Tianjin, China
[2]Institut Langevin, Paris, France
wcrtju01@tju.edu.cn, coder.chen.shi.cheng@gmail.com,
shengguorui@outlook.com, pierre.roussel@espci.fr, denby@ieee.org

ABSTRACT

A 3-dimensional convolutional neural network is trained on unlabeled ultrasound video to predict an upcoming tongue image from previous ones. The network obtains results superior to those of simpler predictors and provides a starting point for exploiting the higher-level representation of the tongue learned by the system in a variety of applications in speech research. This work is believed to be the first application of convolutional neural networks to unlabeled ultrasound video for the purpose of predicting tongue movement.

Index Terms— convolutional neural networks, motion detection, unsupervised learning, speech production, ultrasound, tongue, silent speech interface

1. INTRODUCTION

In the past several years, deep convolutional neural networks, CNNs, have become the state of the art technique for object recognition. Such architectures are believed to encode high-level representations of objects by leveraging massive image databases and powerful graphical processing unit (GPU) hardware [1]. A crucial difficulty of the method is obtaining sufficient labeled data to train the networks. Recently, building high-level representations by training networks to predict future configurations from past behavior in video data has been proposed as a means of eliminating the expensive labeling step [2-5].

Ultrasound imaging of the tongue has been used for many years in research on speech production and pathology [6]; nevertheless, reliable extraction of high-level features from ultrasound data – a labeled contour for example – remains a challenge. In the present article, we investigate for the first time whether CNNs can be used to predict tongue motion from unlabeled ultrasound video sequences, with the aim of building a high-level representation of tongue dynamics for applications in speech research.

2. RELATION TO PRIOR WORK

The applicability of deep learning techniques to raw ultrasound data was confirmed in [7][8], where a deep belief network performed static tongue contour extraction, as well as in [9][10], where CNNs were applied to supervised classification of phonetic targets in the context of a silent speech interface [11].

As CNNs convolve input data with fixed kernels, they were initially thought a poor fit to motion detection, where data-data correlation techniques have been the norm. In [12], however, CNNs nevertheless gave the best performance on supervised optical flow learning tasks. A difficulty discovered in applications of CNNs to prediction of future video frames is the tendency of networks trained on intensity-based mean squared error (MSE) objective functions to output "fuzzy" versions of past images rather than concentrating on movement. The use of novel objective functions [3] and vector-quantization of the image space [2] have been proposed to address this difficulty – unfortunately at a cost of increased computational complexity. A 3DCNN [13] includes the time dimension explicitly by exploring stacks of consecutive input images with 3-dimensional kernels. This architecture gave excellent results on the related task of supervised human gesture recognition [13][14]. The present article describes the first use of a CNN – in this case 3DCNN – for predicting tongue motion in unlabeled ultrasound data, also with good results.

The image datasets used are outlined in the next section. Section 4 details the CNN architecture and training procedure, while results and discussion are presented in section 5. Conclusions and future perspectives appear in section 6.

3. DATASETS AND ENHANCEMENTS

3.1. Datasets

Data were recorded as 320x240 pixel sagittal ultrasound tongue images using an acquisition helmet that stabilizes a 4-8 MHz, 128-element, microconvex ultrasound probe

beneath the speaker's chin. Images were re-sized to 96x96 pixels. The experiments are based upon two principal datasets, detailed below.

1. "WSJ0" data, 60 frames per second, derived from the Silent Speech Challenge containing over 700,000 TIMIT training images and some 35,000 WSJ0 test images, as described in [15];
2. "TJU" data, 30 frames per second, produced at Tianjin University by a volunteer reading a simple training passage (9900 images) and test passage (4800 images). A region of interest, ROI, containing the tongue was selected before resizing.

3.2. Snake contour extraction on the TJU dataset

Tongue contour visibility in the WSJ0 data was mediocre, but the speaker in the TJU data imaged very well, allowing the extraction of an estimated contour for each TJU image with a snake algorithm [16]. These were used both to devise a complementary experiment, as described below, and as a tool to test performance.

4. 3DCNN ARCHITECTURE AND TRAINING

A number of architectures, including CNN variants and recurrent neural networks, RNNs, were tested before settling on the 3DCNN described below. Many solutions worked well on "sharp" input features such as a snake contour, but only the 3DCNN gave satisfactory performance on the noisy, diffuse features of raw images – perhaps because the time-stacked aspect of the 3DCNN allows for noise suppression via averaging. The architecture developed is illustrated in figure 1. It consists of 6 layers, with feature map multiplicities of 1-16-32-64-32-16-1. The input to the network for all experiments was of size 96x96x8, and the output 96x96. Each of the first 3 layers performs a 3D convolution, max pooling, and batch normalization, while the last 3 layers perform up-sampling, convolution, and batch normalization (except last layer).

Three experiments were designed and carried out:

1. "WSJ0": Eight consecutive WSJ0 images were used to predict the 9th WSJ0 image, using an MSE objective function between the prediction and actual next image.
2. "TJU": Eight consecutive TJU images were used to predict the 9th TJU image, using an MSE objective function between prediction and actual next image.
3. "Cross": Eight consecutive TJU images were used to predict the snake contour of the 9th TJU image, using an MSE objective function between an image of the predicted next snake and an image of the actual next snake; compared images contain only the snake, without the ultrasound background image.

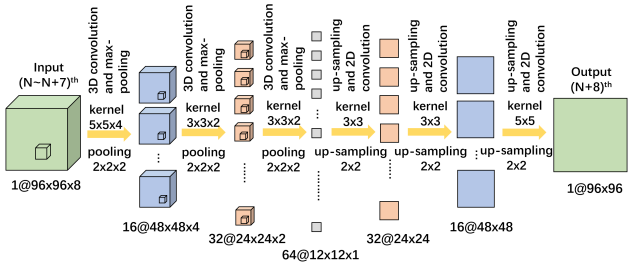

Figure 1. Structure of the 3DCNN used in the tests.

5. RESULTS AND DISCUSSION

5.1 MSE Performance

In table I, the MSE performance of the 3DCNN 9th image predictor is compared to that of three other predictors: the average of the preceding 8 images; the 8th image alone; and a linear predictor based on the previous 8 images. In all experiments, the 3DCNN gives superior results (note that MSE values cannot not be compared *across* columns, due to the different data types). Since it is not possible to output a snake using Average, 8th image, or Linear predictors, the MSE marked with a * in the Cross column is that between the *snakes* of the 8th and 9th images, as a rough comparison (see also the discussion of Cross results in section 5.3).

Table I. Mean MSE on 3 datasets for different predictors

	WSJ0	TJU	Cross
Average	39.2	73.6	-
8th image	31.0	40.0	279.5*
Linear	27.9	38.1	-
3DCNN	**21.7**	**32.6**	**154.9**

*MSE between snakes of 8th and 9th images

It is instructive to examine the time evolution of the MSE per image over several seconds, see Figure 2. A careful inspection of the data confirmed that the "peaks" in the plots correspond to high tongue velocities, and the "valleys" to more stable configurations. The 3DCNN predictor is thus best at predicting rapid tongue movement, while the linear or 8th image predictors are often better in static situations – also explaining the behavior at the beginning and end of the WSJ0 plot where the speaker assumed a "rest" tongue position before and after each sentence. Poorer performance on stable configurations is understood as the inability of the 3DCNN to model time-correlated speckle noise, whereas the simpler predictors contain this intrinsically. That this is indeed the correct interpretation, and that 3DCNN predictions remain of good quality in these regions, was verified through careful inspection. The two 3DCNN "poor performance bumps" near images 60 and 240 in the TJU data, on the other hand, were found to correspond to fixed-position ultrasound artifacts (floor of the mouth, palatal trace, etc.). This

phenomenon was not observed in the WSJ0 data, presumably due to the bigger training set, nor in the Cross data, where the snake images used ignored such artifacts. Finally, the WSJ0 figure shows that the Average predictor introduces a significant delay, and is not useful here.

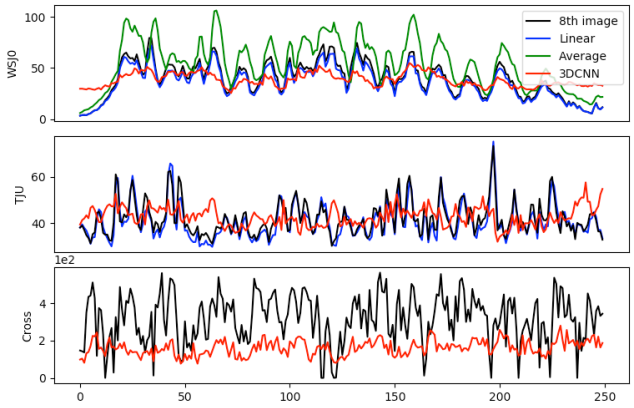

Figure 2. MSE of different predictors for the WSJ0 Challenge Data (top); TJU Data (center); and Cross Data (bottom) over a 250 image sequence (about 4 seconds). The 3DCNN predictor, in red, gives better results on all three datasets. N.B.: For clarity, the Average predictor (green) does not appear in TJU and Cross plots. Also, in the Cross plot, the Linear predictor is not included, and the 8th image MSE refers to the snake of the 8th image, see text.

5.2 Validating motion detection

In order to validate that the good MSE performance of 3DCNN is indeed related to movement prediction, and not the result of "blurring", as mentioned in Section 2, two tests were performed. The first was to compare an overlay of the 8th and 9th WSJ0 images, to an overlay of the 8th WSJ0 image and the 3DCNN prediction of the 9th WSJ0 image. The result appears in Figure 3, where we note that the overlap of a green and a red pixel produces a yellow one. The slightly upward-shifted red image, corresponding in the left panel, to the true 9th image, and on the right, to its prediction, is clearly visible with respect to the green 8th image – particularly in the tongue contour area near the tops of the images. This sequence was selected from a high tongue velocity region, as the shift is often difficult to discern.

The second test devised to show that good 3DCNN results arise from movement prediction and not simple "blurring" was performed on the TJU data making use of the extracted snake contours. A "snake update" here consists of moving each snake point to the center of the brightest 3-pixel square area within a defined "valid" region. To perform the test, the snake was updated on 8 consecutive ultrasound images, and, for the 9th update, either on the 9th true ultrasound image, or on the 3DCNN *prediction* of the 9th ultrasound image. The results are shown for two example cases, in the upper and lower panels of Figure 4. In each example, the green snake of the 8th ultrasound image is compared, on the left, to the red snake of the 9th true image, and on the right, to the red snake of the 9th predicted image. It is seen that the 3DCNN algorithm can indeed follow tongue movement in the TJU data with an error level of only a pixel or two.

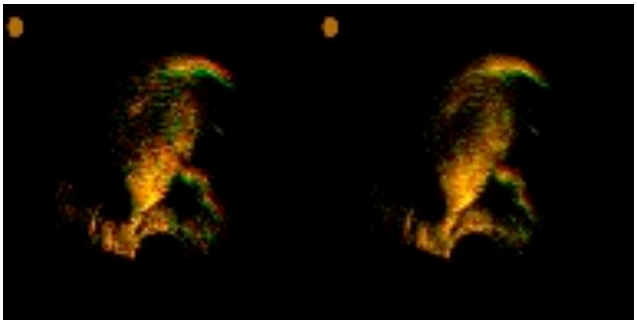

Figure 3. Image sequence from WSJ0 Challenge Data. Left: overlay of 8th image (green) and 9th image (red); Right: overlay of 8th image (green) and 3DCNN prediction of 9th image (red). Tongue movement reproduced by 3DCNN prediction is visible in the tongue contour near the top of the image. N.B. The overlap of a red and a green pixel appears yellow.

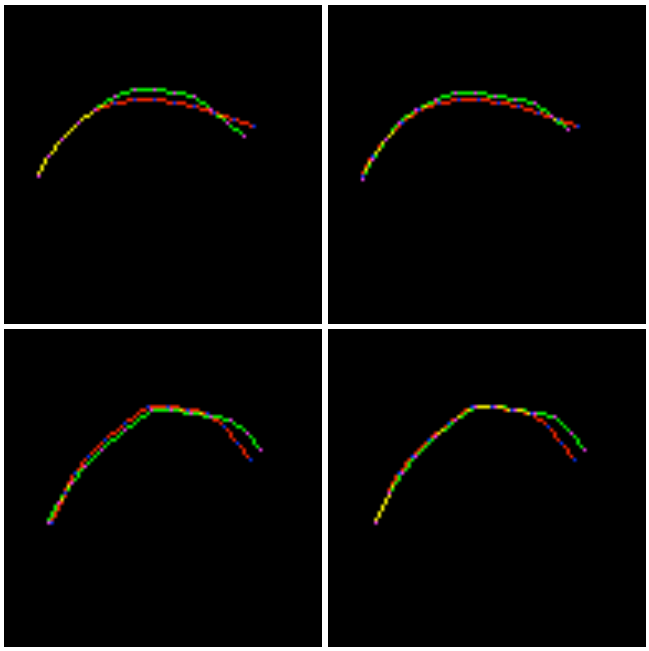

Figure 4: TJU Data. Top left: overlay of snake of 8th image (green) and snake of 9th image (red); Top right: overlay of snake of 8th image (green) and snake of 9th image prediction (red). Bottom left and right: same overlays for another sequence.

5.3 Relevance of the Cross dataset

The MSE performance and the preceding tests demonstrate that the 3DCNN is indeed able to predict tongue motion in real time ultrasound video. This however is, of course, not the goal of the exercise. If the technique has indeed enabled

the construction of an internal representation that models image content and dynamics [3] – of the tongue in this case – then it ought to be possible to exploit this representation to perform some useful task of importance to speech processing. The Cross dataset was created to enable a first look at this hypothesis.

The Cross dataset training, as described in section 4, is similar in some respects to the contour finding carried out in [7][8], with the key difference that here, the prediction refers to a *future* contour. The question posed is whether the 3DCNN can isolate the dynamics of a precise, acoustically relevant feature – in this case the sagittal tongue contour – using only unprocessed ultrasound images as input, with the caveat that, contrary to the WSJ0 and TJU cases, labeled contours are in fact a necessary element this time. The technique works extremely well, as can be seen in the videos available at [17][18], where true and predicted snakes atop ultrasound images, as well as overlays of the two obtained snakes, respectively, are exhibited. The Mean Sum of Distances [19], MSD, between the 3DCNN prediction and the 9th image snake was measured to be 1.1 pixels, corresponding to 0.4 mm for these data. Outtakes from the videos for three examples appear in Figure 5. The third example shows a case in which the true and predicted snakes are rather different; interestingly, the predicted contour may actually be the more correct one. The success of the Cross data experiment is thus a promising first step towards creating a higher-level model of tongue dynamics that may be useful in a variety of concrete applications in speech research.

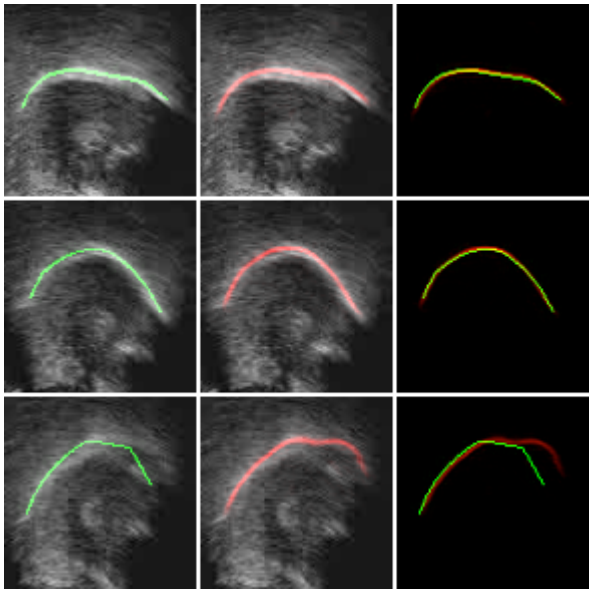

Figure 5. Snake of 9th image (left, in green); snake of predicted 9th image (center, in red); and overlay of the two curves (right); for three example tongue contour shapes.

6. CONCLUSIONS AND PERSPECTIVES

A 3DCNN has been shown capable of predicting future frames in raw ultrasound video with pixel level accuracy. When trained to predict instead a tongue contour, results suggest that the system possesses an internal representation of tongue dynamics that could provide useful input for subsequent speech research tasks. Future work will include more powerful GPU hardware; studies of alternate objective functions; experiments with vector quantization of the image space (as in [2]); and, finally, a test of the method as a front-end for a silent speech recognition system.

7. ACKNOWLEDGMENTS

Partial funding for this work was provided by the China Ministry of Education "985 Foundation" via grant number 060-0903071001. The authors wish to thank the reviewers for many useful suggestions for improving the article.

8. REFERENCES

[1] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks", *Proceedings of Advances in Neural Information Processing Systems NIPS*, pp. 1097–1105, Curran Associates, Inc., 2012.

[2] M. Ranzato, A. Szlam, J. Bruna, M. Mathieu, R. Collobert, S. Chopra, "Video (language) modeling: a baseline for generative models of natural videos", arXiv: 1412.6604v4 [cs.LG] 21 Apr 2015.

[3] M. Mathieu, C. Couprie, Y. Le Cun, "Deep multi-scale video prediction beyond mean square error", *Proceedings of International Conference on Learning Representations ICLR*, San Juan, Puerto Rico, May 2016.

[4] C. Finn, I. Goodfellow, S. Levine, "Unsupervised learning for physical interaction through video prediction", *Proceedings of Advances in Neural Information Processing Systems NIPS*, Barcelona, Spain, December 2016.

[5] A. Canziani, E. Culurciello, "CortexNet: a generic network family for robust visual temporal representations", arXiv: 1706.02735v2 [cs.CV] 14 Jun 2017.

[6] Y.S. Akgul, C. Kambhamettu, M. Stone, "Automatic extraction and tracking of the tongue contours", *IEEE Transactions on Medical Imaging*, 18, 1035-1045, 1999.

[7] I. Fasel, J. Berry, "Deep belief networks for real-time extraction of tongue contours from ultrasound during speech", *Proceedings of 20th International Conference on Pattern Recognition ICPR*, Istanbul, Turkey, August 2010.

[8] A. Jaumard-Hakoun, K. Xu, P. Roussel-Ragot, G. Dreyfus, B. Denby "Tongue contour extraction from ultrasound images based on deep neural network", *Proceedings of International Congress of Phonetic Sciences ICPhS*, Glasgow, UK, August 2015.

[9] E. Tatulli, T. Hueber, "Feature extraction using multimodal convolution neural networks for visual speech recognition", *Proceedings of ICASSP2017*, New Orleans, USA, March 2017.

[10] K. Xu, P. Roussel, T. Gábor Csapó, B. Denby, "Convolutional neural network-based automatic classification of midsagittal tongue gestural targets using B-mode ultrasound images", *The*

Journal of the Acoustical Society of America 141, EL531, 2017.

[11] B. Denby, T. Schultz, K. Honda, T. Hueber, J.M. Gilbert, J.S. Brumberg, "Silent Speech Interfaces", *Speech Communication*, vol. 52, pp. 270-287, 2010.

[12] P. Fischer, A. Dosovitskiy, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, T. Brox, "FlowNet: learning optical flow with convolutional networks", *Proceedings of International Conference on Computer Vision ICCV*, Santiago, Chile, December 2015.

[13] S. Ji, W. Xu, M. Yang, K. Yu, "3D Convolutional neural networks for human action recognition", *Proceedings of 27th International Conference on machine Learning ICML*, Haifa, Israel, 2010.

[14] P. Molchanov, S. Gupta, K. Kim, J. Kautz, "Hand gesture recognition with 3D convolutional neural networks", *Proceedings of the Workshop on Computer Vision and Pattern Recognition CVPR*, Boston, USA, June 2015.

[15] B. Denby, T. Hueber, J. Cai, P. Roussel, L. Crevier-Buchman, S. Manitsaris, G. Chollet, M. Stone, C. Pillot, "The Silent Speech Challenge Archive", 2013:
online: https://ftp.espci.fr/pub/sigma/.

[16] M. Kass, A. Witkin, D. Terzopoulos, "Snakes: Active contour models", *International Journal of Computer Vision*, 1.4, 321-331, 1988.

[17] https://youtu.be/oXP4DyMbwC4

[18] https://youtu.be/E25BDTOtC08

[19] M. Li, R. Kambhamettu, M. Stone, "Automatic Contour Tracking in Ultrasound Images", *Clinical Linguistics and Phonetics*, 19, 545-554, 2005.