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

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

[1] Tianjin University, Tianjin, China

[2] Institut langevin, ESPCI ParisTech, PSL Research University, Paris, France

Wechat

Github

Introduction:

- A 3-dimensional convolutional neural network is trained on unlabeled ultrasound videos to predict an upcoming tongue image from previous ones.
- The network obtains results superior to those of simpler predictors
- A starting point for exploiting the higher-level representation of the tongue learned by the system in a variety of applications in speech research.

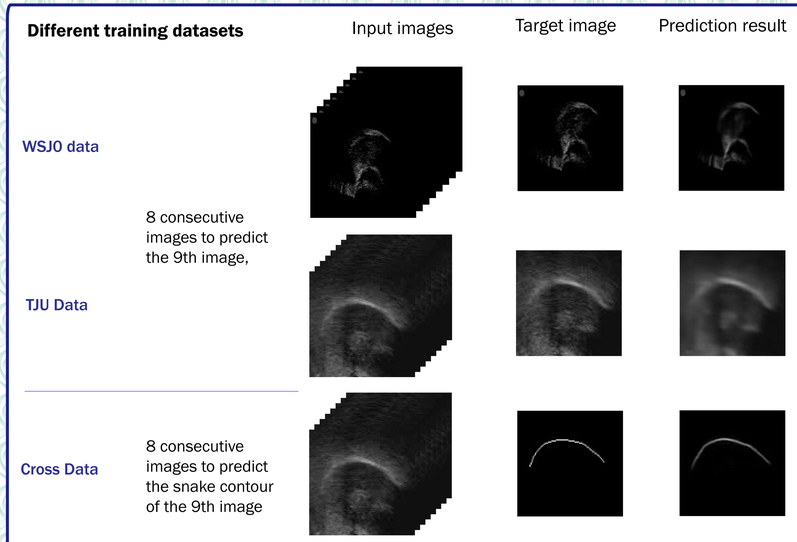
Why doing this:

- To use *Unsupervised Learning* to extract features from tongue images for speech recognition.
- To prove simple time regression problem can be handled well without RNN.

What did we do:

- Prediction of ultrasound video frame (WSJ0 & TJU data)
- Prediction of utrasound tongue contour (Cross data)
- Evaluation metric design:
  - calculating the loss (MSE) between different predictors and the target image
  - overlapping different prediction results and the target image
  - validating motion detection method to detect the contour change in the predicted frame

Different training datasets



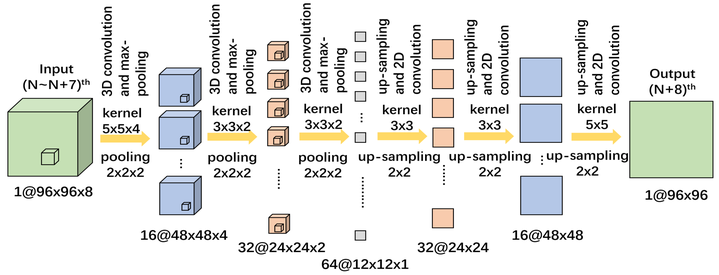
Why 3D CNN:

- To take time dimension information into consideration.
- To preserve better spatial (2D) information than RNN.
- The time-stacked aspect of the 3DCNN allows for noise averaging.

Why do these tests:

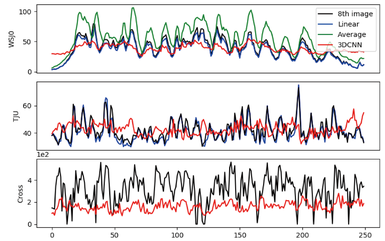
- To show that we can do better than simple predictors.
- To prove that the predicted frames are actually moved not simply blurred.
- To visualize the result.

Structure of 3DCNN



MSE of 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**



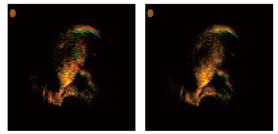
- In all experiments, the 3DCNN gives superior results.
- 3D-CNN works best on Cross data

- The 3DCNN predictor, in red, gives better results on all three datasets.

- The 3DCNN predictor works better on the moving part.

Overlaying 8th image and 9th/predicted image

WSJ0 data

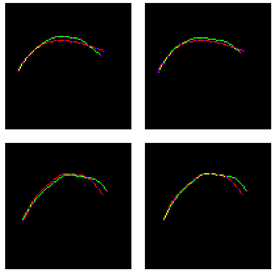


Left: overlay of the 8th image (green) and 9th image (red)
Right: overlay of the 8th image (green) and 3DCNN predicted 9th image (red).

- The red tongue contour is **above** the green one in both images.
- This means the predicted image actually moved rather than just blurring.

Motion detection from contour change

TJU data



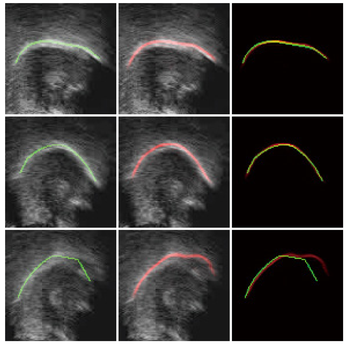
Left: comparison of the contours of the 8th image (green) and the 9th image (red);
Right: comparison of the contours of the 8th image (green) and the predicted 9th image (red).

How did we do this:

- Initialize the video with the same 12 points.
- Extract the contours from the real images and the predicted ones.
- If the predicted image's snake has the same position as the real one, then we learnt the 'movement'.

Prediction of the tongue contour

Cross data



Contour of the 9th image (left, in green);
Predicted contour of the 9th image (center, in red);
Overlay of the two contours (right).

The technique works extremely well, as can be seen in the videos available (scan the QR code)

video for the two left columns

video for the right column

- The Mean Sum of Distances, MSD, between the 3DCNN prediction and the 9th image snake was only **1.1 pixels**, corresponding to **0.4 mm**.