ObamaNet: Photo-realistic lip-sync from text

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present **ObamaNet**, the first architecture that generates both audio and synchronized photo-realistic lip-sync videos from any new text. Contrary to other published lip-sync approaches, ours is only composed of fully trainable neural modules and does not rely on any traditional computer graphics methods. More precisely, we use three main modules: a text-to-speech network based on **Char2Wav**, a time-delayed LSTM to generate mouth-keypoints synced to the audio, and a network based on **Pix2Pix** to generate the video frames conditioned on the keypoints.

1 Introduction

Data driven approaches for generating images have recently surpassed traditional computer graphics methods (see for example Isola et al. (2016)). In parallel, there has been significant progress in speech synthesis (see for example Sotelo et al. (2017)). In this work, we show that we can combine some of these recently developed models to generate artificial videos of a person reading aloud an arbitrary text. Our model can be trained on any set of close shot videos of a person speaking, along with the corresponding transcript. The result is a system that generates speech from an arbitrary text and modifies accordingly the mouth area of one of the videos so that it looks natural and realistic. A video example can be found there: `http://ritheshkumar.com/obamanet`

Although we showcase the method on Barack Obama because his videos are commonly used to benchmark lip-sync methods (see for example Suwajanakorn et al. (2017)), our approach can be used to generate videos of anyone provided the data availability.

2 Related Work

Recently, there have been important advances in the generation of photo-realistic videos (Thies et al., 2016). In particular Karras et al. (2017) have tried to generate facial animations based on audio. The work by Suwajanakorn et al. (2017) is the closest to ours, yet we have two important differences. First, we replace the computer vision model with a neural network. Second, we add a text-to-speech synthesizer in order to have a full text-to-video system.

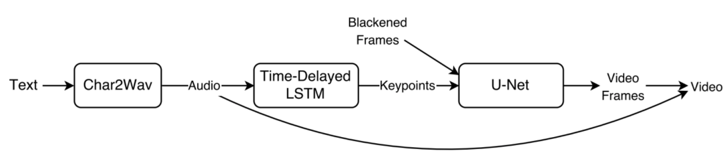

Figure 1: Flow diagram of our generation system.

3 Model Description

3.1 Text-to-speech system

We use the Char2Wav architecture (Sotelo et al., 2017) to generate the speech using the input text. We train the speech synthesis system using the audio extracted from the videos, along with their corresponding transcripts.

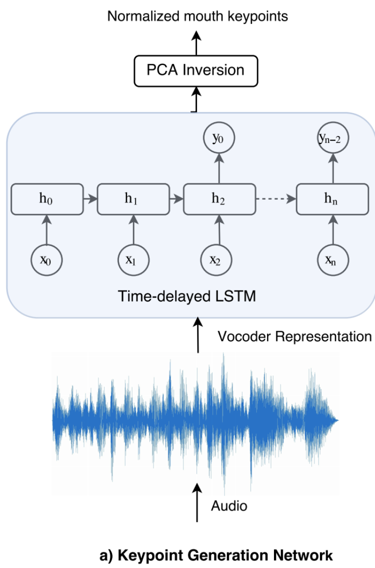

Figure 2: Keypoint Generation Network

3.2 Keypoint generation

This module predicts the representation of the mouth shape, given the audio as input. We use spectral features to represent the audio. To compute the mouth-shape representation, we use mouth keypoints extracted from the face, and normalize the points to be invariant to image size, face location, face rotation and face size. Normalization is crucial in the pipeline, as it makes the key-point generation compatible with any target video. We then apply PCA over the normalized mouth key-points to reduce the dimension and to decorrelate the features. We only use the most prominent principal components as the representation for mouth shape. Further details can be found in (Put section label here) Supplementary Material: Data Processing.

For the network, we adopt the same architecture as Suwajanakorn et al. (2017). We use a time-delayed LSTM network to predict the mouth shape representation given the audio features as input.

3.3 Video generation

Our motivation behind the choice of method to perform video generation is the recent success of pix2pix (Isola et al. (2016)) as a general-purpose solution for image-to-image translation problems. This task falls within our purview, as our objective here is to translate an input face image with cropped mouth area, to an output image with in-painted mouth area, conditioned on the mouth shape representation.

To avoid explicit conditioning of mouth shape representation in the U-Net architecture, we implicitly condition by drawing an outline of mouth on the input cropped image. The network learns to leverage this outline to condition the generation of the mouth in the output.

We noticed that the keypoints generated by the recurrent network are consistent across time without abrupt changes. This allowed us to perform video generation in parallel, by synthesizing each frame in the video independently across time, given the conditioning information of the mouth keypoints. We did not need any explicit mechanism to maintain temporal consistency in the generated frames of the video.

We train this network only using L1-loss in pixel-space and found that this objective is sufficient to learn the in-painting of the mouth and doesn't require the extra GAN objective as originally proposed in pix2pix by Isola et al. (2016).

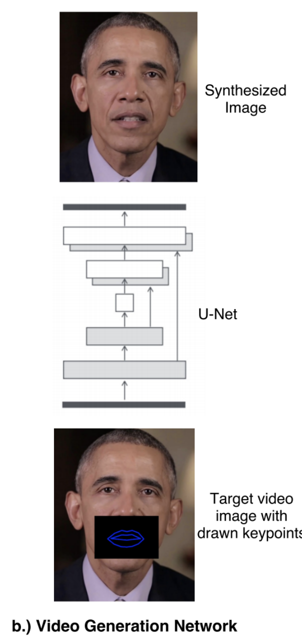

Figure 3: Video Generation Network

4 Supplementary Material

4.1 Dataset

We showcase our approach on videos of ex-President Barack Obama, similar to Suwajanakorn et al. (2017). We used 17 hours of video footage from his 300 weekly presidential addresses, which have the benefit to frame the president in a relatively controlled environment, with the subject in the center of the camera.

4.2 Data Processing

Text to speech We extract the audio from the videos and convert it to 16KHz. We extract vocoder frames from the audio using the WORLD vocoder, and use the transcript associated with the video to train the text-to-speech system.

Figure 4: The 68 facial keypoints

Keypoint Generation The data required for the keypoint generation component is a representation of audio for input, and a representation of mouth shape for the output.

To compute the mouth shape representation, we extract 68 facial keypoints from each frame of the video. We used the publicly available dlib facial landmark detector to detect the 68 keypoints from the image. Sample annotations performed by the detector are shown in Figure 3.

These keypoints are highly dependent on the face location, face size, in-plane and out-of-plane face rotation. These variances are due to varying zoom-levels of the camera, distance between camera and speaker, and the natural head-motion of the speaker. In an effort to remove these variances, we first mean-normalize the 68 keypoints with the center of the mouth. This converts the 68 keypoints into vectors originating from the center of the mouth, thereby making it invariant to the face location.

To remove the in-plane rotation caused due to head motion, we project the keypoints into a horizontal axis using rotation of axes.

We make the keypoints invariant to face size, by dividing the keypoints by the norm of the 68 vectors from the center of the mouth, which serves as an approximation of face size.

Finally, we apply PCA to de-correlate the 20 normalized keypoints (40-D vector). We noticed that the first 5 PCA-coefficients capture >98% variability in the data.

Video Generation The data required for this component is image pairs, where the input face image is cropped around the mouth area and annotated with the mouth outline and the output image is the complete face.

For this task, We extract 1 image per second of video for all 300 videos, extracting keypoints from these images using the dlib facial landmark detector. We crop the mouth area from each image using a bounding box around the mouth keypoints, and the mouth outline is drawn with keypoints 49-68 using OpenCV. Figure 4 shows a sample input / output pair.

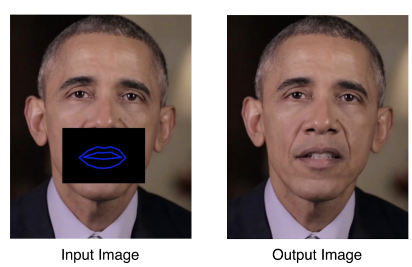

Figure 5: Sample input-output pair for the in-painting network

An important aspect of the video generation process is to denormalize the generated keypoints from the previous stage of the pipeline, with the mouth location, size and rotation parameters of the target video. This ensures that the rendered mouth is visually compatible with the face in the target video.

References

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016.

Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Trans. Graph.*, 36(4):94:1–94:12, July 2017. ISSN 0730-0301. doi: 10.1145/3072959.3073658. URL `http://doi.acm.org/10.1145/3072959.3073658`.

Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio. Char2wav: End-to-end speech synthesis. 2017.

Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4):95, 2017.

J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2016.