FaceSyncNet: A Deep Learning-Based Approach for Non-Linear Synchronization of Facial Performance Videos

Yoonjae Cho*  Dohyeong Kim*  Edwin Truman  Jean-Charles Bazin
KAIST
{yoonjaecho, dohyeong, edwin.truman, bazinjc}@kaist.ac.kr

Abstract

Given a pair of facial performance videos, we present a deep learning-based approach that can automatically return a synchronized version of these videos. Traditional methods require precise facial landmark tracking and/or clean audio, and thus are sensitive to tracking inaccuracies and audio noise. To alleviate these issues, our approach leverages large-scale video datasets along with their associated audio tracks and trains a deep learning network to learn the audio descriptors of a given video frame. We then use these descriptors to compute the similarity between video frames in a cost matrix and compute a low-cost non-linear synchronization path. Both quantitative and qualitative evaluations have shown that our approach outperforms existing state-of-the-art methods.

1. Introduction

This paper is dedicated to the synchronization of facial performance videos, that is the temporal alignment of different video takes of a scene performed by actors. Synchronization of facial performance videos is a fundamental step for various tasks in computer vision and computer graphics such as video morphing [11], face replacement [5] and generating novel performances of actors [13].

When multiple cameras are simultaneously used to record different viewpoints of a scene, synchronization boils down to a constant time offset and can be easily achieved using timecode information or camera triggers. In contrast, we investigate the case of video takes that are shot at different times, for example when an actress repeats the same scene possibly with different emotions, facial expressions and timing. Due to the complex non-linear local variations in timing and speed during facial performances, such videos cannot be synchronized by a simple constant time offset or uniform temporal scaling.

A way to synchronize videos is to manually align them using video editing programs such as Adobe Premiere. While this is feasible for videos with a constant time offset, it is extremely challenging, tedious and time consuming for non-linear video takes. Some methods have been proposed to synchronize videos automatically based on the visual content. However, they are designed for large-scale scene changes [24, 17, 7] or require precise facial landmark detection and tracking [13, 5].

Our approach is motivated by the observation that the challenging visual task of video synchronization can be greatly simplified with audio information. For example, audio provides precious cues for speech videos (such as facial performances), audio synchronization is well studied [4, 23] and audio can inherently overcome several visual challenges (facial landmark, lighting conditions, tracking, head/camera motion, etc). Therefore some audio-based methods have been developed for video synchronization [13, 4, 23]. However they require clean speech audio of actor’s video takes (which typically needs an audio-controlled environment) and thus, are vulnerable to audio noise, such as background audio, music and talking.

Inspired by the recent works on audio-visual data [6, 21, 19, 15, 8, 14, 30, 20], we propose FaceSyncNet: a deep learning-based approach for predicting the audio features of actor’s facial performances in the context of video synchronization. For this, we leverage large-scale video datasets along with their associated audio tracks and train a convolutional neural network (CNN) to output the (clean) audio descriptors of a given video frame. We then use these descriptors to compute the similarity between video frames in a cost matrix and compute a low-cost non-linear synchronization path. Experimental results show that our approach outperforms existing state-of-the-art methods, especially when the audio is corrupted with noise.

To the best of our knowledge, this paper presents the first deep learning-based approach for non-linear synchronization of facial performance videos, which constitutes the key contribution of our work. Overall, in contrast to existing methods, our approach does not require the extraction and tracking of facial landmarks, does not assume clean audio-controlled recording environment, and does not need complicated head/camera motion separation.

*denotes joint first authorship with equal contribution.
2. Related Work

2.1. Hand-crafted Visual Frame Descriptor

Various hand-crafted frame descriptors have been studied for video synchronization [17, 7, 24]. While these methods work best with camera ego-motion and large-scale scene changes, they cannot capture the subtle changes of facial expressions [13]. Therefore some methods have been specifically proposed for facial performances. For example, Yang et al. [27] employed the facial expression coefficients of a morphable face model but their method can only handle large changes of facial expression. Dale et al. [5] used the velocity of the mouth landmarks in the context of face replacement and FaceDirector [13] used both facial landmarks and audio information. However, these methods are sensitive to landmark tracking inaccuracies, require sophisticated methods to separate head/camera motions and/or assume clean audio.

2.2. Metric Learning

Motivated by the recent deep learning achievements, some metric learning-based approaches with image triplets have been proposed to compute video frame similarity for synchronization tasks [26, 18]. For example, Wieschollek et al. [26] synchronized outdoor driving scene videos taken at different times and weather conditions. Sermanet et al. [18] learned representations from multiple viewpoints videos in the context of imitation learning of human motions by a robot. However, in our early experiments, we observed that metric learning does not provide reliable synchronization results for facial performance videos (see results in supplementary material).

2.3. Audio Visual Information

Our approach is also related to recent works exploiting the correlation between video and audio information [6, 21, 19, 15, 8, 14, 30, 20]. Among them, some methods have been proposed to estimate the audio of an image/video for various contexts, such as the sound of an object being hit by a stick [14], separating a single speech signal from a mixture of sounds such as other speakers and background noise [8], recovering sound from the subtle movement of an object surface vibrating [6] and retrieving sounds corresponding to a query image (e.g. car sound from a car image) [20]. Our work also takes advantage of the correlation between visual and audio information but is dedicated to the context of facial performance video synchronization. In our work, we demonstrate that audio descriptors can be learnt for actor’s video takes and be successfully applied for non-linear synchronization of facial performance videos.

3. Proposed Approach

3.1. Learning Audio Descriptor

As discussed, we propose a deep learning-based approach to learn audio descriptors for video synchronization. Among several existing audio descriptors, we selected Mel-frequency cepstrum coefficients (MFCC). Several studies have demonstrated that MFCC is a useful representation for temporal alignment of audio signals [4, 23] as well as for audio analysis, description and retrieval [12, 16, 25]. MFCC can be represented as a 13 dimension vector.

Leveraging large-scale video datasets along with their associated audio tracks (see Sec. 4.1), we train a deep learning network to predict the MFCC descriptor of a given video frame. The input of our network is a frame of a facial performance video (whole frame, cropped face or cropped mouth by MTCNN face detection model [29]), and the output is a predicted 13-dimensional MFCC audio descriptor. We computed the “ground truth” MFCC descriptor from the audio signals associated to every video frame. For loss, we used the L2 distance between the normalized “ground truth” and predicted MFCC descriptors [13, 20]. For network architecture, we used InceptionResNetV1 model [22] pretrained with VGGFace2 [3]. It is worthwhile to note that our approach just needs videos along with their associated audio tracks for CNN training and thus can use any unlabeled videos in the wild.

3.2. Computing Cost Matrix and Synchronization Path

We now explain how to compute the synchronization of a pair of input videos $v_1$ and $v_2$. For each frame of the videos, we employ our CNN to predict their MFCC audio descriptor. We then use these audio descriptors to compute a cost matrix $C$ whose cell entries represent the dissimilarity between two video frames:

$$C_{i,j} = \|f_1(i) - f_2(j)\|_2$$

where $f_1(i)$ is the predicted audio descriptor of the $i$-th frame of video $v_1$, $f_2(j)$ is the predicted audio descriptor of the $j$-th frame of video $v_2$, and $\|f_1(i) - f_2(j)\|_2$ refers to the $L2$ distance between the audio descriptors (Sec. 3.1). Finally, we compute the lowest cost path in the cost matrix using dynamic programming [2]. This non-linear synchronization path represents the frame mapping between the two videos, i.e. which frame of $v_1$ corresponds to which frame of $v_2$, with which we synchronize the two videos.

4. Experiments

4.1. Implementation and Training

All the experiments were conducted on a computer equipped with Intel Xeon-E5-1650v4 CPU, dual GTX
Figure 1. Representative evolution of the estimated synchronization path (in blue) and the associated cost matrix along the epochs. It shows our path gets closer to the ground truth path (in red) and the cost matrix resembles more the “ground truth” MFCC cost matrix (right most).

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean (ms)</th>
<th>Median (ms)</th>
<th>Max (ms)</th>
<th>Fréchet [1] (ms)</th>
<th>Hausdorff [9] (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours (full frame)</td>
<td>416 (12.5f)</td>
<td>283 (8.5f)</td>
<td>1300 (39.0f)</td>
<td>526 (13.8f)</td>
<td>1223 (36.7f)</td>
</tr>
<tr>
<td>Ours (face cropped)</td>
<td>136 (4.1f)</td>
<td>56 (1.7f)</td>
<td>640 (19.2f)</td>
<td>383 (11.5f)</td>
<td>496 (14.9f)</td>
</tr>
<tr>
<td>Ours (mouth cropped)</td>
<td>432 (12.9f)</td>
<td>380 (11.4f)</td>
<td>1296 (38.9f)</td>
<td>966 (29.0f)</td>
<td>1166 (35.0f)</td>
</tr>
<tr>
<td>Video Face Replacement [5]</td>
<td>266 (7.7f)</td>
<td>100 (2.7f)</td>
<td>1233 (37.4f)</td>
<td>533 (16.5f)</td>
<td>966 (29.1f)</td>
</tr>
<tr>
<td>FaceDirector [13]</td>
<td>400 (12.0f)</td>
<td>366 (11.2f)</td>
<td>1000 (30.4f)</td>
<td>500 (15.2f)</td>
<td>933 (27.7f)</td>
</tr>
<tr>
<td>MFCC</td>
<td>800 (23.5f)</td>
<td>800 (23.6f)</td>
<td>1600 (47.9f)</td>
<td>833 (24.6f)</td>
<td>1433 (42.8f)</td>
</tr>
</tbody>
</table>

Table 1. Quantitative results on our video dataset with different error metrics. The metric unit is expressed in milliseconds (ms) and number of frames (f) and refers to the time difference between the ground truth synchronization path and the one obtained by different methods.

4.2. Quantitative Evaluation

One way to evaluate our approach could be to measure the accuracy of the MFCC descriptors predicted by our network. However, a certain score of the MFCC metric would not be easy to interpret in the context of video synchronization. Instead, we directly evaluate the synchronization quality by measuring the temporal distance between the ground truth synchronization path and ours. To this end, we need different video takes of a scene. For this, we captured our own videos and added realistic background noise, such as cafe ambient sound. Our video dataset is composed of 20 videos (10 pairs) of scenes performed by 10 actors/actresses at different locations and with different emotions. In the following, “ground truth” refers to the MFCC cost matrix computed on the clean audio (i.e. without added noise) and the associated synchronization path.

Fig. 1 shows that along the training epochs, our predicted path is getting closer to the ground truth path and our cost matrix has similar appearance with the ground truth MFCC cost matrix (see supplementary material for more examples). We investigated various path distance and curve similarity metrics: mean, median, max, Fréchet [1] and Hausdorff [9] distance (see details in supplementary material). For example, the mean distance is defined as the normalized frame distance between two paths:

$$d(P, G)_{mean} = \frac{1}{n} \sum_{t=1}^{n} |P(t) - G(t)|$$  \hspace{1cm} (2)

where $P(t)$ and $G(t)$ respectively correspond to the $t$-th frame index on the predicted and ground truth paths, and $n$ is the number of video frames.

In Table 1, we compare the metrics obtained by several variations of our proposed approach (full frame, cropped face and cropped mouth), the previous methods [5, 13] and MFCC. It shows that the cropped face input for our approach provides better metrics than the whole frame and cropped mouth, and outperforms all the other methods. Moreover, Fig. 3 shows that our approach satisfies the perceptual audio-video synchronization error threshold (185.19 ms, 5.5 frame distance at 30 fps video [28]) in 78% of the test dataset video frames.

4.3. Qualitative Evaluation

We conducted a user study to compare the performance of our approach with previous works [5, 13] and MFCC. 10
Figure 2. Our method can handle various types of videos such as different camera viewpoints (a, e), different emotions (b), wearing accessory such as hat and glasses (c, e, f), head-mounted camera (d), hand-held camera (e) and different depths of field (f).

Figure 3. Evaluation of different video synchronization methods. X axis is the error distance between the ground truth and predicted paths (in milliseconds and in frames). Y axis represents the cumulative distribution within the error. The vertical dashed line is the perceptual error threshold [28].

Figure 4. Distribution of the user study grades. people participated and we showed 40 pairs of synchronized videos side by side (10 video pairs, and for each of them, we applied 4 methods, so a total of 40 pairs), which represents a total of 400 votes. The participants were asked to respond to the perception statement “This pair of videos is temporally aligned” by choosing a response from a five-point Likert scale: strongly agree (5), agree (4), neither agree nor disagree (3), disagree (2) or strongly disagree (1).

The distribution of the scores is available in Fig. 4. First, it shows that our approach outperforms existing methods. Second, the participants strongly agreed in 54.4% of our results, and a total of 94.4% had a score equal to or higher than 4 (agree). This demonstrates that our approach provides perceptually satisfying synchronization results for facial performance videos.

4.4. Additional Results

Fig. 2 shows additional synchronization results on various types of videos. Our approach provides reliable results for videos with different viewpoints (Fig. 2.a, e), different emotions (Fig. 2.b), people wearing accessory such as hat and glasses (Fig. 2.c, e, f), head-mount camera (Fig. 2.d) and cameras with different depths of field (Fig. 2.f).

Limitations The experiments demonstrated that our approach provides satisfying results for non-linear video synchronization of facial performance videos. However, since deep learning performance depends on the training dataset, our network performance is limited to our current training dataset, i.e. AVSpeech. Therefore our current network cannot handle videos with extreme difference of camera viewpoints or lighting for example (see Fig. 5). This can be alleviated with training dataset covering these cases. Our current approach processes each frame independently. In future work, time series network such as RNN and LSTM could be used to take advantage of temporal information.

5. Conclusion

We have presented a novel deep learning-based approach for non-linear synchronization of facial performance videos. Given a video frame, we predict its audio descriptor and use it to compute frame similarity, and in turn a synchronization path. The experiments demonstrated that our approach provides satisfying synchronization results and outperforms existing state-of-the-art methods.

Acknowledgement This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2018-0-01077, Development of Deep Learning AI Based Face Replacement Technology for the Construction of Realistic Digital Actors in Modern Film CG/VFX Production).
References


