ROBUST AUDIOVISUAL LIVENESS DETECTION FOR BIOMETRIC AUTHENTICATION USING DEEP JOINT EMBEDDING AND DYNAMIC TIME WARPING

Amit Aides[1,2]*, David Dov*[1,2]*, Hagai Aronowitz*[1]

[1]IBM Research AI, Israel
[2] Viterbi Faculty of Electrical Engineering, Technion - Israel Institute of Technology, Israel
{amitaid, hagaia}@il.ibm.com
davidd@tx.technion.ac.il

ABSTRACT

We address the problem of liveness detection in audiovisual recordings for preventing spoofing attacks in biometric authentication systems. We assume that liveness is detected from a recording of a speaker saying a predefined phrase and that another recording of the same phrase is a priori available, a setting, which is common in text-dependent authentication systems. We propose to measure liveness by comparing between alignments of audio and video to the a priori recorded sequence using dynamic time warping. The alignments are computed in a joint feature space to which audio and video are embedded using deep convolutional neural networks. We investigate the robustness of the proposed algorithm across datasets by training and testing it on different datasets. Experimental results demonstrate that the proposed algorithm generalizes well across datasets providing improved performance compared to competing methods.

Index Terms: Cross-database generalizability, Text dependent speaker recognition, Spoofing countermeasure, Audiovisual synchronization, Deep-Learning

1. INTRODUCTION

Audiovisual biometrics is an appealing technology for applications such as mobile authentication, in which the goal is to identify a person based on recordings of, e.g., his voice and face. Nowadays, even the most basic mobile devices are equipped with a camera and a microphone which enable the capture of audiovisual content. Furthermore, the fusion of two highly uncorrelated biometrics such as face and voice has the potential for accurate and robust authentication.

However, the abundance of mobile devices and social media increases the risk of playback attacks that target the face [1] and voice modalities. For example, an authentication system, which is merely based on face recognition, will wrongly identify a person based on a still image of his face, presented by an attacker. Therefore, the task of *Liveness* detection, which we address in this paper, is becoming a key factor in real life authentication systems. In the absence of special hardware like infra-red light sources or depth cameras, it is hard to ensure liveness. This is why in some methods the user is prompted to express certain facial expressions, or repeat random pass-phrases, which may lead to accuracy loss (Table 4 in [2]).

Following [3], we propose to exploit synchrony between audio and video recordings for liveness detection. The appeal of this approach stems from the difficulty in spoofing both modalities (audio and video) simultaneously, as opposed to spoofing each modality independently. Specifically, we consider a setting, in which a speaker is required to say a certain phrase during an enrollment phase. Then, during authentication, the speaker repeats the same phrase and the two recordings are compared for liveness detection. We refer to this problem settings as text dependent and note that is was previously addressed in [4], where the authors suggested using Dynamic Time Warping (DTW) [5] to find two temporal alignments which map between test and enrollment sequences, one for each modality. Then, the difference between the audio and video mappings is used as a measure of liveness such that the more the mappings are similar to each other, the higher the confidences that the tested recording is authentic. In this context, we also note methods e.g., [6, 7], which do not rely on enrollment and directly measure the synchrony between the tested audio and video sequences. The main limitation of these methods, whether they utilize an enrollment phrase or not, stems from the large difference between the audio and video modalities in terms of dimensionality, value ranges, and temporal dynamics. Therefore, it is not clear how to define their correspondence.

We propose to alleviate this problem by embedding audio and video into a joint feature space. Such approach gains increasing interest in recent years since it allows exploiting complex relations between modalities. Related methods may be roughly divided into non parametric approaches such as kernel based geometric methods [8–10], and parametric methods such as deep neural networks [11]. In this paper we take the latter utilizing a Deep Learning (DL) framework, presented in [12]. The use of DL approaches has led in recent years to unprecedented improvement in the accuracy of many audiovisual tasks such as face and speaker recognition [13–15] and lip-reading [16–18]. In the context of spoofing countermeasures, a major challenge in the development of deep learning methods is the lack of suitably large datasets. Till such databases are available, an important criterion by which such algorithms should be evaluated is their cross-database applicability [19]. Specifically, the performance of such methods may significantly deteriorate when trained and evaluated on different datasets, as we show in this paper.

In this paper, we present an algorithm for liveness detection, which is based on utilizing two convolutional neural networks, one for each modality. These networks are specifically designed to embed audio and video into a joint domain. Once embedded, the audio and video of authentic recordings should appear similar to each other. In the new domain, we compute two DTW alignments between the enrollment and authentication recordings, one for the audio stream and a second for the video stream. Then, we exploit the similarity between the audio and the video based alignments as a measure for liveness. We demonstrate that audio and video alignments between the enrollment and authentication recordings are indeed more similar to each other in the embedded domain for authentic videos, compared to domains which are designed separately for audio and video. Then, we show that the proposed algorithm out-

performs the methods presented in [12] and [4] in terms of Equal Error Rate (EER). Finally, we show that the proposed algorithm is highly invariant to the training domain so that it provides significantly higher detection scores compared to [12] when these algorithms are tested on different datasets than they were trained on.

The rest of the paper is organized as follows. The proposed algorithm is presented in Section 2. In Section 3, we describe the different datasets and the experimental method and present the improved performance of the proposed algorithm. Finally, we discuss future research directions and conclude our work in Section 4.

2. ALGORITHM

2.1. Pre-processing

We process both audio and video streams in consecutive frames similarly to [12]. The audio is represented by the Mel Frequency Cepstral Coefficients (MFCCs) [20], which represent the spectrum of the signal in a compact form and are widely used in audio processing applications. We use 12 coefficient extracted at a frame rate of 100 fps excluding the first coefficient, which corresponds to the energy of the signal. A Voice Activity Detector (VAD) is applied prior to the extraction of the MFCCs to detect and remove leading and trailing silences from both audio and video streams. In the video stream, we crop a Region of Interest (ROI) of size 120×120 pixels around the lips of the speaker using a face detection algorithm. Then, 20 consecutive MFCC frames and 5 corresponding video frames are stacked together representing $\approx$ 200 ms long sequence. Care is taken so that the samples span the same temporal chunk of the video.

2.2. Embedding Audio and Video Into a Joint Domain

We follow the method presented in [12] to embed the audiovisual stacks into a joint feature space. The embedding is performed by feeding the audio and the video stacks into two corresponding convolutional neural networks (CNN). The audio stack is viewed by the CNN as a single channel, two-dimensional image of size 12×20. Similarly, the video is considered as an image with 5 channels (features) such that the tensor to the network has the size of $120 \times 120 \times 5$. Both networks comprise 5 convolutional layers with Rectified Linear Unit (ReLU) used as a non-linearity followed by two fully connected layers. We denote the outputs of the networks by $\boldsymbol{\alpha}_k \in \mathbb{R}^L$ and $\boldsymbol{\nu}_k \in \mathbb{R}^L$ for the audio and the video, respectively, where k denotes the index of the time frame, and L is set to 256. We refer the reader to [12] for more details on the architectures of the CNNs and on the construction of the audiovisual features.

The networks are trained using the following loss function, which we denote by L:

$$L = \sum_k y_k \left\|\boldsymbol{\alpha}_k - \boldsymbol{\nu}_k\right\| + (1 - y_k)\max\left(C - \left\|\boldsymbol{\alpha}_k - \boldsymbol{\nu}_k\right\|, 0\right), \tag{1}$$

where y_k is an indicator which equals one for authentic pairs and zero for spoofed, and C is a constant value. By design, the loss function encourages the networks to learn embeddings, for which authentic audio and video pairs are mapped close to each other, while spoofed pairs are embedded distantly from each other. The embedding of audio and video into the joint feature domain is illustrated in fig. 1.

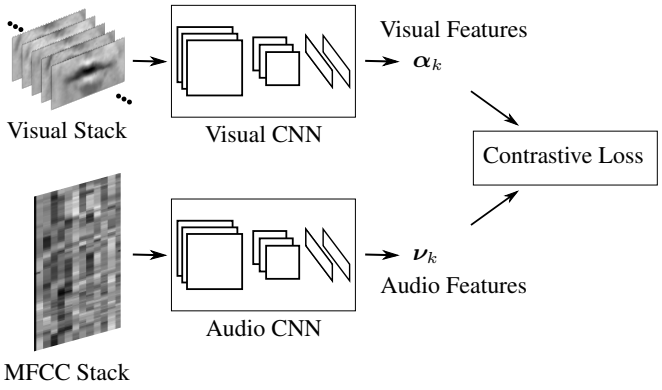


Fig. 1: *Embedding audio and video into a joint domain.*

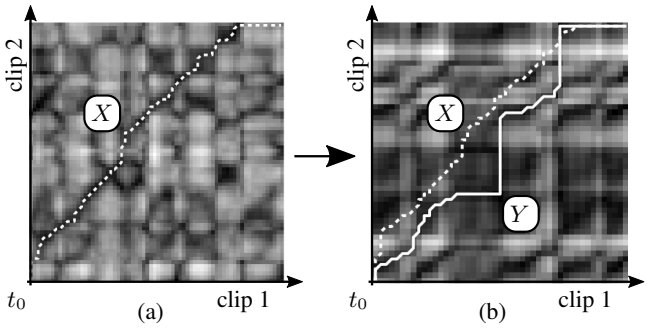


Fig. 2: *(a) A matrix of distances between the enrollment and tested audio feature sequences, with the corresponding alignment path X marked by a dashed line. (b) A matrix of distances between the enrollment and tested video feature sequences. The projected audio alignment path is marked by a dashed line. The visual alignment path Y, as calculated by the DTW method, is marked by a solid line.*

2.3. Liveness measure

Similarly to [4], we propose a liveness measure based on DTW. Contrary to [4] our novel liveness measure is based on, first, embedding both the enrollment and the tested sequences into the joint audio-visual domain using the CNNs. Then we construct two pairwise distance matrices between the enrollment and the tested sequences, one for each modality such that the i, k-th element in each matrix represents the Euclidean distance between the i-th enrollment frame and the k-th tested frame. Finally, we exploit the DTW method to find a temporal alignment path X which maps between the audio sequences of the test and the enrollment clips. Similarly, we find a second temporal alignment path Y that maps between the respective video sequences as is illustrated in fig. 2.

The proposed measure of liveness, which we denote by S_{DTW}, is based on the Hausdorff distance, and is given by:

$$S_{\mathrm{DTW}} = \max\{\frac{1}{|X|}\sum_{x\in X}\inf_{y\in Y}\left\|x-y\right\|, \frac{1}{|Y|}\sum_{y\in Y}\inf_{x\in X}\left\|x-y\right\|\}, \tag{2}$$

where a smaller S_{DTW} refers to a higher confidence that the tested recording is authentic. The Hausdorff distance measures how far subsets of a metric space, in our case the alignment paths, are from each other. It finds the maximal distance between each point in one subset to its closest neighbor in the other set. Here we use the aver-

(a) *Still image of BBC News videos (taken from [12])*

(b) *Still images taken from our mobile dataset*

Fig. 3: *Examples of datasets referenced in this work*

age distance instead since we empirically found it performing better for liveness detection than other metrics we tested.

The proposed algorithm for liveness detection allows for a meaningful comparison both between the modalities, using the joint audiovisual embedding and within the modalities using the dynamic time warping algorithm. Accordingly, the similarity between the alignment paths X and Y reliably indicates on liveness as we further discuss in Section 3.3.

3. EXPERIMENTAL RESULTS

3.1. Datasets

The networks we use in our experiment were trained on a large dataset compiled of BBC News programs. According to the authors (the dataset is not publicly available) it is ≈ 700 hours long, and contains audiovisual samples of a large number of different people [12]. In Fig. 3a, we present an example of a single video frame taken from the BBC News videos dataset.

As the target dataset, we use the same "mobile" dataset used in [4]. This dataset is made of clips collected using an iPad-2 and an iPhone-5. The dataset comprises two or three recorded sessions for each of the 41 subjects on each device. During the sessions each subject repeats the following phrases three times: *my voice is my password*; and *please verify me with the number*. The dataset comprises 1906 recordings with average length of 1.5s. Figure 3b shows examples of video frames taken from the target dataset.

After the preprocessing stage, which includes the cropping of the mouth ROI, the recordings in both datasets comprise a single front-facing speaker. Nonetheless, the datasets significantly differ from each other in the recording conditions. The BBC dataset was captured in good lighting and sound conditions at a distance of up to several meters between speaker and camera using professional recording equipment. The "mobile" dataset was recorded with smartphones and tablets held at arm's length. This setting significantly degrades the quality of the audio signal and lighting. Also, the short distance between speaker and camera results in considerable distortion of facial proportions.

3.2. Experimental Setup

We design a spoofing scenario, in which we use all the audiovisual recordings in the "mobile" dataset as positive (authentic) examples. For each authentic recording, we create a corresponding spoofed recording by replacing the authentic video with a different recording of the same speaker, while keeping the original audio. Accordingly, the spoofed audiovisual recording comprises audio and video, which does not correspond to each other. This challenging setting resembles a scenario, in which an attacker is trying to spoof an authentication system using an audio recording of a certain speaker saying the correct passphrase and a different video recording of the same speaker. In the experiments, we consider the existence of one or three enrollment recordings. When three recordings are available, the integrated liveness score is the minimum of the three scores.

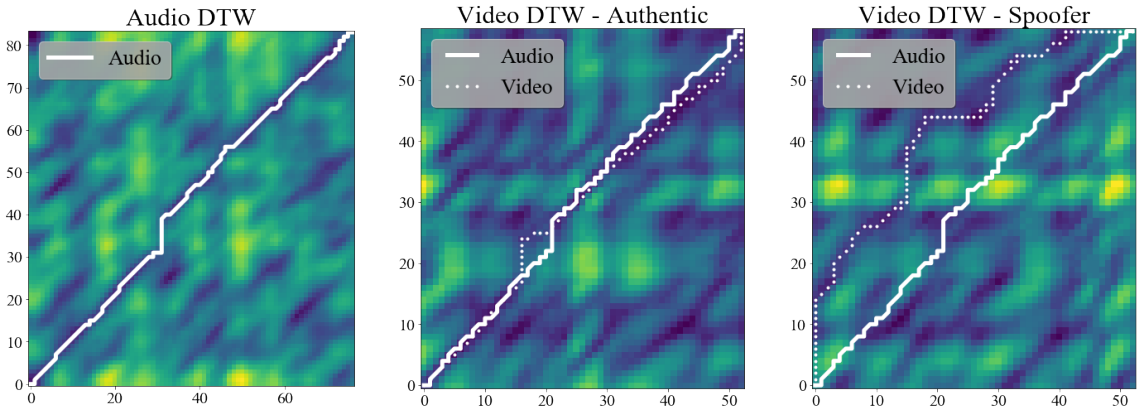
We demonstrate the performance of the proposed algorithm in two experiments. First, we train the proposed algorithm on the "mobile" dataset such that the training data does not comprise the same speakers as the tested data. Then, we study the ability of the proposed algorithm to generalize across datasets by using the original networks from [12] that were trained on the BBC dataset to produce the audio and video embeddings, α_k and ν_k, respectively. In our experiments, we compare the proposed algorithm to the methods presented in [4] and [12]. While the method in [4] is designed for a similar task as in this study, the method in [12] does not assume the availability of enrollment recordings. Specifically, the authors suggest measuring the synchronization between audio and video simply using the l_2 norm between the embeddings α_k and ν_k such that the smaller the norm the higher is the synchronization. Yet, we find it reasonable to compare between the methods to demonstrate the contribution of the DTW measure especially for the generalizability of the proposed algorithm across datasets.

3.3. Results

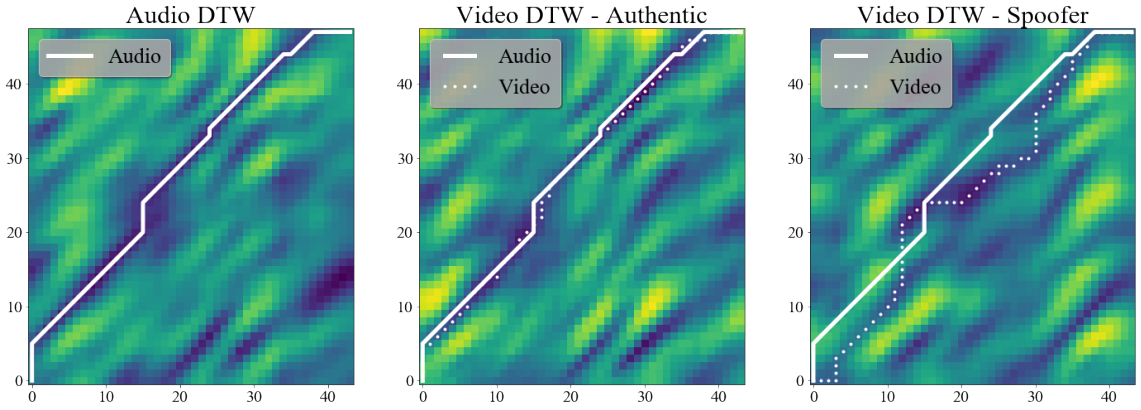
We present in fig. 4 the distance matrices of the audio, authentic video and spoofed video obtained by the method in [4], as well as by the proposed method. The high similarity between the distance matrices of the authentic pair, obtained by the proposed algorithm, demonstrates that the CNNs indeed satisfactorily map audio and video into a joint feature domain. As a result, the corresponding DTW based alignment paths, based on the audio and the video, indeed appear more similar to each other than those obtained in [4].

The results of the first experiment are summarized in table 1 in terms of Equal Error Rate (ERR) such that the lower the ERR the better. It can be seen that the proposed method based on three enrollment recordings outperforms the competing methods demonstrating the usefulness of the incorporation of CNN embeddings and the DTW algorithm. Specifically, the audio and the video of both the enrollment and the tested recordings are mapped into a joint domain allowing for a reliable comparison between them, which, in turn, leads to successful liveness detection.

In table 2, we present the results of the second experiment, where we use versions of the CNNs pre-trained on the BBC News dataset. The performance of the method presented in [12], which achieves $> 99\%$ accuracy on the BBC News dataset as is reported in [12], significantly deteriorates, when it is tested on a mismatched dataset. In contrast, the proposed algorithm successfully generalize between the datasets performing significantly better in this scenario.



(a) *The DTW measure applied to embeddings calculated using the network in [4].*



(b) *The DTW measure applied in the joint audiovisual domain.*

Fig. 4: *Visualization of the DTW scoring method*

Table 1: *EERs (in %) for the networks trained on the "mobile" dataset.*

Passphrase	[4]	[12] (l_2)	S_{DTW} (proposed)	S_{DTW} 3 enrolls (proposed)
My voice...	4.1	0.7	1.68	0.72
Please verify...	5.1	3	2.84	0.44
Average	4.6	1.85	2.26	0.58

Table 2: *EERs (in %) for the original networks (cross dataset generalizability).*

Passphrase	[12] (l_2)	S_{DTW} (proposed)	S_{DTW} 3 enrolls (proposed)
My voice...	32.71	2.98	1.3
Please verify...	31.07	4.39	2.49
Average	31.89	3.69	1.9

These results demonstrate the robustness of the DTW algorithm and its usefulness for practical liveness detection.

4. CONCLUSIONS

We have presented an algorithm for liveness detection addressing a challenging setting, where the spoofer comprises an audio recording of the correct passphrase and a different video of the same speaker. The algorithm is based on measuring the similarity between enrollment and tested recordings in terms of DTW alignment paths computed in a joint audio-visual domain, obtained using two convolutional neural networks corresponding to the two modalities. Due to the lack of large anti-spoofing datasets, we have considered a setting, in which the CNNs are trained and tested on different datasets and demonstrated that the proposed algorithm generalizes well across datasets thanks to the similarity measure based on the DTW algorithm. Our plans for future work include the improvement of the joint audio-visual embeddings. Being based on CNNs, the embeddings do not fully capture the temporal dynamics of the audio and the video signals, and specifically for the video CNN, the convolutions are applied in the spatial domain while the consecutive frames are merely considered different channels. Therefore, using deep architecture such as Long Short-term Memory network (LSTM) for obtaining the joint audio-visual representation may further improve liveness detection.

5. REFERENCES

[1] Sean Hollister, "Psa: Your note 8's face unlock can easily be fooled," 2017.

[2] Hagai Aronowitz, Ron Hoory, Jason Pelecanos, and David Nahamoo, "New developments in voice biometrics for user authentication," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2011.

[3] Girija Chetty and Michael Wagner, "Liveness verification in audio-video authentication," in *Proceedings of the 10th Australian International Conference on Speech Science and Technology (SST04)*, 2004, pp. 358–363.

[4] Amit Aides and Hagai Aronowitz, "Text-dependent audiovisual synchrony detection for spoofing detection in mobile person recognition," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 08-12-Sept, pp. 2125–2129, 2016.

[5] Hiroaki Sakoe and Seibi Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," *IEEE transactions on acoustics, speech, and signal processing*, vol. ASSP-26, no. 1, pp. 43–49, 1978.

[6] Hervé Bredin and Gérard Chollet, "Audiovisual speech synchrony measure: Application to biometrics," *Eurasip Journal on Advances in Signal Processing*, vol. 2007, 2007.

[7] Etienne Marcheret, Gerasimos Potamianos, Josef Vopicka, and Vaibhava Goel, "Detecting Audio-Visual Synchrony Using Deep Neural Networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] D. Dov, R. Talmon, and I. Cohen, "Kernel-based sensor fusion with application to audio-visual voice activity detection," *IEEE Transactions on Signal Processing*, vol. 64, no. 24, pp. 6406–6416, Dec 2016.

[9] D. Dov, R. Talmon, and I. Cohen, "Multimodal kernel method for activity detection of sound sources," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1322–1334, 2017.

[10] R. R. Lederman and R. Talmon, "Learning the geometry of common latent variables using alternating-diffusion," *Applied and Computational Harmonic Analysis*, 2015.

[11] I. Ariav, D. Dov, and I. Cohen, "A deep architecture for audio-visual voice activity detection in the presence of transients," *Signal Processing*, vol. 142, pp. 69–74, 2018.

[12] Joon Son Chung and Andrew Zisserman, "Out of time: Automated lip sync in the wild," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10117 LNCS, no. i, pp. 251–263, 2017.

[13] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," *Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 8, 2014.

[14] Yun Lei, Nicolas Scheffer, Luciana Ferrer, and Mitchell McLaren, "A Novel Scheme for Speaker Recognition Using a Phonetically-Aware Deep Neural Network," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014*, pp. 1714–1718, 2014.

[15] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer, "End-to-End Text-Dependent Speaker Verification," , no. Section 3, pp. 3–7, 2015.

[16] Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, and Nando de Freitas, "LipNet: End-to-End Sentence-level Lipreading," pp. 1–12, 2016.

[17] Themos Stafylakis and Georgios Tzimiropoulos, "Combining Residual Networks with LSTMs for Lipreading," 2017.

[18] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Lip Reading Sentences in the Wild," *Eccv*, vol. 9905, pp. 1–17, nov 2016.

[19] Keyurkumar Patel, HuHan, and Anil K. Jain, "Cross-Database Face Antispoofing with Robust Feature Representation," in *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9428, pp. 611–619. 2016.

[20] P. Davis S. and Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Signal Processing*, vol. 28, no. 4, pp. 357 – 366, 1980.