Text-Dependent Audiovisual Synchrony Detection for Spoofing Detection in Mobile Person Recognition

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

[1]IBM Research - Haifa, Israel
[2]Dept of Electrical Engineering,Technion - Israel Institute of Technology, Haifa, Israel
{amitaid, hagaia}@il.ibm.com

Abstract

Liveness detection is an important countermeasure against spoofing attacks on biometric authentication systems. In the context of audiovisual biometrics, synchrony detection is a proposed method for liveness confirmation. This paper presents a novel, text-dependent scheme for checking audiovisual synchronization in a video sequence. We present custom visual features learned using a unique deep learning framework and show that they outperform other commonly used visual features. We tested our system on two testing sets representing realistic spoofing attack approaches. On our mobile dataset of short video clips of people talking, we obtained equal error rates of 0.8% and 2.7% for liveness detection of photos and video attacks, respectively.

Index Terms: text dependent speaker recognition, anti spoofing, audiovisual synchronization, deeply learned audio features

1. Introduction

Audiovisual biometrics is an appealing technology for applications such as mobile authentication. Authentication that uses a video recording of a person saying a short pass-phrase or prompt does not require non-standard hardware. Furthermore, the fusion of two highly uncorrelated biometrics such as face and speaker recognition has the potential for accurate and robust authentication. Aronowitz et al. [1] recently investigated the merits of audiovisual biometrics, obtaining an Equal Error Rate (EER) of $\sim 0.3\%$ for mobile authentication in smartphones and tablets. In the past two years, the use of deep learning has led to even higher accuracy in face recognition [2] and in speaker recognition [3, 4].

Following these recent breakthroughs in accuracy, a considerable amount of focus has shifted to the risk of spoofing. Special sessions in Interspeech [5, 6] and in the TABULA-RASA EU-funded project [7] specifically investigated the risk of spoofing and corresponding countermeasures.

Optimal speaker recognition accuracy requires the use of common pass-phrases such as *my voice is my password* [1] or *OK Google* [4]. Conversely, the use of a prompted pass-phrase (less vulnerable to spoofing), may increase the EER by a factor of five (Table 4 in [8]). User-selected pass-phrases may provide a possible compromise between common pass-phrases and prompted pass-phrases. However, both common and user-selected pass-phrases are vulnerable to playback and audio-splicing attacks.

The face modality is also easily vulnerable to a playback attack. Liveness is hard to ensure unless the user is prompted to express certain facial expressions or poses.

Chetty [9] proposes the use of synchrony detection in audiovisual biometrics. The appeal of this approach in liveness detection stems from the difficulty in spoofing both modalities simultaneously, as opposed to the ease of spoofing each modality independently. Bredin [10] provides a survey of past work, as does the more recent paper by Marcheret [11]. Most works try to find some kind of correlation or correspondence between the time series of audio-based features and the time series of visual-based features, usually focusing on the mouth as Region Of Interest (ROI).

The most commonly used acoustic features are the energy [12] or spectral based MFCC features [13]. Suggestions for visual features include pixel values [12], Discrete Cosine Transform (DCT) coefficients [14], lip-based measurements (height, width, etc.) [9], optical flow [15],and more recently the scattering-transform [11] and deep learning-based features [11].

The methods used to estimate the correspondence between audio and visual-based features are: mutual-information [12], Canonical Correlation Analysis [13], Coinertia Analysis [16], maximally informative projections [17], Hidden Markov Models [14], Generalized Bimodal Linear Prediction [18], a time-delay neural network (TDNN) [19] or a deep neural network [11].

Bregler and Konig [20] showed that the mutual information between the audio and video streams was maximal when the lag between audio and video data was approximately 120ms (in their dataset). Furthermore, the lags are highly context-sensitive. This finding motivates the non-trivialness of the task. Other works [21, 22] reported similar findings.

This paper shows a different approach. We are interested in a text-dependent authentication scenario in which the enrollment pass-phrase is identical to the verification pass-phrase (pass-phrase is either common or user-selected). We exploit this attribute and deviate from the generic synchrony detection framework.

We define a new task that we name text-dependent audiovisual synchrony detection. This task assumes the existence of an enrollment audiovisual recording from the target person (the person we are trying to authenticate), with the same spoken text (pass-phrase) as in the test recording.

We base the motivation for our method on the following: for every synchronized test recording, there exists a natural alignment between the enrollment and test recordings. We can find the alignment for each modality separately, and the similarity between the two modality-dependent alignments gives a strong indication regarding the synchrony of the test recording. In principle, we can avoid low-level audiovisual processing; that is, we never compare audio low-level features with visual low-level features.

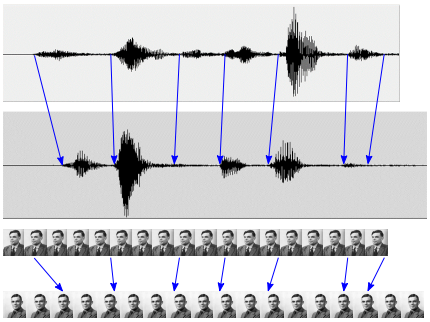

Figure 1: *Illustration of the proposed method for text-dependent synchrony detection. Alignments are computed independently for the audio and visual modalities, and are expected to be similar for synchronized recordings*

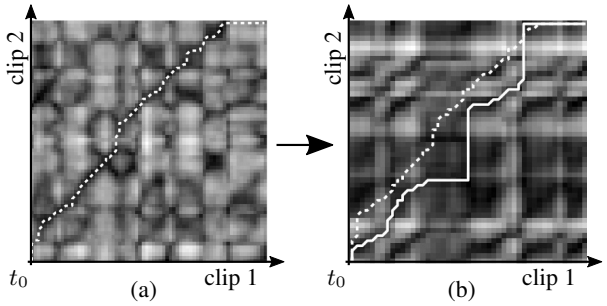

Figure 2: *(a) Distance Matrix of the audio features, with the corresponding alignment path marked by a dashed line. (b) Distance matrix of the visual features. The projected audio alignment path is marked by a dashed line. The visual alignment path, as calculated by the DTW method, is marked by a solid line.*

We experimented with our proposed approach on a multi-subject, multi-session audiovisual dataset recorded with smartphones and tablets held at arms length. This setting significantly degrades the quality of the audio signal. We observed that the MFCC features are suitable for estimating accurate alignment for the audio modality. The alignments obtained for the visual modality using standard features such as Histogram of Gradients (HOG), were much less accurate. Therefore, we introduce a novel, visual-based features using deep learning, with an objective function that targets finding an accurate alignment.

The rest of this paper is organized as follows: We present our proposed text-dependent audiovisual synchrony detection scheme in Section 2. We then provide an overview of the feature extraction methods in Section 3. We present our dataset, experiments, and results in Section 4. Finally, we conclude our work in Section 5.

2. Text-dependent Audiovisual Synchrony Detection

This paper focuses on the following scenario: for a given test audiovisual recording (clip), there exists an enrollment clip from the target person (the real person we are trying to authenticate), with the same text spoken as in the test clip.

For a synchronized test clip, we claim that a single modality-independent temporal alignment exists between the enrollment and test clips (Figure 1). We can estimate modality-dependent alignments by processing each modality independently, and measure the similarity between the modality-dependent alignments. We then use the similarity degree as an indication for audiovisual synchrony of the test recording.

The framework we propose has the following advantages. First, we can avoid low-level audiovisual processing (the temporal alignments are per modality). Second, the context-dependent lagging reported by Bregler [20], which degrades the standard techniques reviewed in section 1, does not affect our method. Lastly, our proposed framework generally does not require any audiovisual training data. In practice, to achieve the best results, we use audiovisual training data to train visual-based features, as shown in section 3.

2.1. Algorithm

Given enrollment and test clips, we first apply a voice activity detector (VAD) on the audio streams to locate and remove leading and trailing silences from both audio and visual streams. The audio streams are then divided into evenly-spaced overlapping frames (50 frames per second (fps)), and each frame is represented by a feature vector. Correspondingly, the visual streams are composed of sequences of frames (30 fps). Each frame is also represented by a feature vector.

Let $a_{1,1}, \ldots, a_{1,m_1}$ and $a_{2,1}, \ldots, a_{2,m_2}$ denote the audio-based feature sequences for enrollment and test clips, respectively. Let $v_{1,1}, \ldots, v_{1,n_1}$ and $v_{2,1}, \ldots, v_{2,n_2}$ denote the corresponding visual-based feature sequences. We use the Dynamic Time Warp (DTW) [23] method to find a temporal alignment $\alpha : [1, \ldots, m_1] \rightarrow [1, \ldots, m_2]$, which maps audio sequence $a_{1,1}, \ldots, a_{1,m_1}$ into $a_{2,1}, \ldots, a_{2,m_2}$, and a temporal alignment $\mu : [1, \ldots, n_1] \rightarrow [1, \ldots, n_2]$, which maps the respective visual sequences. The DTW algorithm outputs both the optimal temporal alignment S_a and a corresponding score S_v. The DTW score is the integral of the corresponding weight matrix along the alignment path.

We re-sample DTW alignment α to match the frame rate of the visual stream (setting now m_1 to n_1 and m_2 to n_2). We thus create a path that represents the audio-induced alignment in the visual domain (Figure 2. We integrate this path on the visual distance matrix, D_v, to get the score of the audio-induced visual alignment $S_{\tilde{v}}$. Next we define the synchrony score S_{sync}, as the difference between the score of the audio-induced alignment in the visual domain and the score obtained using unconstrained DTW in the visual modality,

$$S_{sync} = S_{\tilde{v}} - S_v, \tag{1}$$

where a smaller S_{sync} refers to better synchronization.

3. Features for Synchrony Detection

This section describes the features we use to parameterize the audio and visual streams.

3.1. Audio-based Features: MFCCs

The audio is parameterized by low-level spectral-based Mel scale Cepstral Coefficients (MFCCs) [24], which are standard in speech processing. MFCCs are extracted at a frame rate of 50 fps, 20 coefficients per frame.

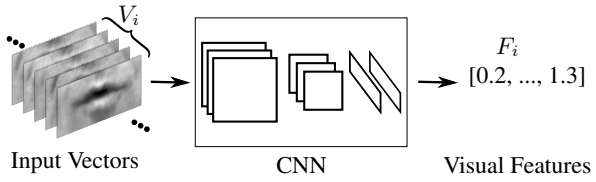

Figure 3: *Illustration of the proposed deep learning-based visual features*

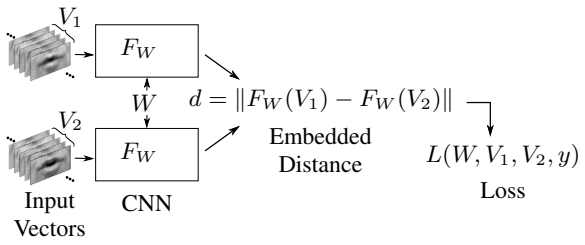

Figure 4: *Illustration of the proposed Siamese network for training the deep visual features for text-dependent synchrony detection*

3.2. Visual Stream Pre-processing

We used the Viola-Jones face detector [25] to first detect the face in each video frame. Next we used ASM (Active Shape Model) [26, 27] to locate the lips, and crop a 30×50 pixel region around the lips. We use only the 30×50 crops for all further processing (section 3.3 and section 3.4). Although we have to establish the facial pose in each frame, state-of-the-art algorithms [27] enable near real-time processing of the video stream.

3.3. Visual-based Features: HOG

HOG is a shape descriptor that has been successfully applied to human detection and face recognition. We compute it as follows: 1) each sample is evenly divided into 16×16 cells; 2) for each cell, we calculate a histogram of 8 gradient orientation bins (in $0 - 2\pi$) (see details of HOG in [28]).

3.4. Visual-based Features: Deep Learning-based

Figure 3 illustrates the deep learning network architecture. The network has three convolution layers, followed by two fully connected layers. Rectified linear units are applied after each layer, except for the last fully connected layer. The input to the Convolutional Neural Network (CNN) is a stack of 5 consecutive 30×50 mouth crops. The motivation for stacking frames in time is to capture lip movement.

We used a Siamese architecture [29] (Figure 4) to train the network. In each iteration, the network is given a pair of lip stacks, V_1 and V_2 and a label y, which is either 0 if the lips differ and 1 if the lips match. The loss function L

$$L = \frac{1}{2N} \sum_{i=1}^{N} y_i d_i^2 + (1 - y_i) \max(\delta - d_i, 0)^2, \qquad (2)$$

where y_i and d_i are the label and Euclidean distance for the i-th lip stack pair respectively, and δ is a predefined margin. The selection of positive pairs of lip stacks for training is as follows:

1. Select a pair of positive clips (same person and text).

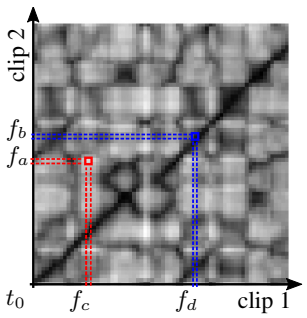

Figure 5: *MFCC distance matrix of two clips. f_a and f_c represent a negative pair while fb and fd represent a positive pair.*

2. Extract MFCC features for each clip.
3. Find the optimal DTW path for the MFCC features.
4. Enumerate pairs of MFCC frames along the DTW path, whose distance is below a predetermined threshold.
5. For each such pair of MFCC frames find a pair of visual frames in the corresponding clips, closest in time to the MFCC frames.
6. For each such pair of visual frames, form a positive training sample using the corresponding lip stacks.

Negative training sample selection is similar. However, instead of sampling along the DTW path, the pairs of frames are sampled at the maximal pairwise distance measure between MFCC frames.

This selection scheme of positive and negative pairs encourages the network to produce visual features that mimic the correspondence between MFCC features. Figure 6 illustrates the selection process.

4. Experiments

4.1. Dataset

We recorded the dataset using an iPad-2 and an iPhone-5. Each of the 41 subjects had two or three recorded sessions on each device. A session includes a subject repeating the following phrases three times: *my voice is my password* and *please verify me with the number*. We used 5-fold cross-validation to train the deep visual features. That is, for each experiment we used 80% of the data for training, and the remaining 20% for evaluation (subjects in the training and the evaluation subsets are disjoint). The average net length of each phrase is 1.5s.

4.2. Training Setup of the Deep Visual Features

The training set was constructed as follows. For each subject and each pass-phrase, we selected all pairs of clips (12915, including cross-device pairs). For each such pair we found an optimal alignment using DTW on the audio (MFCC) stream. We then extracted 60 positive samples (pairs of audio and visual frames) along the DTW alignment, and 60 negative samples taken randomly off the DTW alignment. The total number of samples was 1.5M.

We used 10% of the data set during training to evaluate the learning progress. Again, we made sure that the subjects in the training and evaluation subsets were disjoint.

Figure 6: *Example of an image used for creating the static image video attack (here the face is pixelated for privacy)*

4.3. Photo Spoofing Attack Scenario

The first spoofing scenario we tested simulates an attacker using an audio of the subject saying the correct pass-phrase, and using a static image of the subject as a visual attack. To create the dataset we used one static image per subject (Figure 6. We held each image in front of a camera, and took a video while slightly moving the image to mimic 'liveness'. We created 25 such videos corresponding to 25 subjects from the smartphone dataset.

To create the testing set, we selected all pairs of clips for each subject (of the 25 subjects subset) and each pass-phrase. These served as the positive pairs. We matched a negative pair for each positive pair by replacing the visual stream of the second clip with the static image video of the corresponding subject.

4.4. Live Video Spoofing Attack Scenario

The second spoofing scenario is more challenging. Similar to the training setup, we selected all pairs of clips (4208 pairs, same-device only) for each subject and each pass-phrase. The selected pairs served as positive pairs. For each positive pair we created a corresponding negative pair by replacing the visual stream of the second clip with a different visual stream. We used a different pass-phrase from the original pair of clips.

4.5. Results

Table 1 shows the results for the photo spoofing attack scenario, using deep learning-based visual features. On average, we obtained an EER of 1.55% for a single clip enrollment, and 0.75% for 3-clip enrollment.

Table 1: *EERs (in %) for the photo spoofing attack using deep learning-based visual features*

Pass-phrase	Enrollment: 1 clip	Enrollment 3 clips
My voice...	1.6	0.6
Please verify...	1.5	0.9
Average	1.55	0.75

Table 2 shows the results for the live video spoofing attack scenario using deep learning-based visual features. On average, we obtained an EER of 4.6% for single clip enrollment, and 2.7% for 3-clip enrollment. .

Table 3 shows the results for the live video spoofing attack scenario using HOG features instead of deep learning-based features. The error rates are ~ 3 times higher compared to using deep learning-based features.

Table 2: *EERs (in %) for the live video spoofing attack using deep learning-based visual features*

Pass-phrase	Enrollment: 1 clip	Enrollment 3 clips
My voice...	4.1	2.4
Please verify...	5.1	3.0
Average	4.6	2.7

Table 3: *EERs (in %) for the live video spoofing attack using HOG-based visual features*

Pass-phrase	Enrollment: 1 clip	Enrollment 3 clips
My voice...	17.4	12.4
Please verify...	22.1	17.5
Average	19.8	15.0

5. Conclusions and Future Work

We address the problem of liveness detection by introducing a novel, text-dependent audiovisual synchrony detection scheme.

Our work assumes the availability of enrollment clips for speaker and face recognizers. We further assume that the same pass-phrase is used for both enrollment and authentication. Exploiting these assumptions eliminates the need to actually compare the audio and visual-based low-level features. Furthermore, the availability of enrollment clips implies that the algorithm can be fine-tuned to different use cases, and thus enhance its robustness.

We evaluated our method on two spoofing scenarios. The first is a photo attack, in which we obtained an EER of less than 1%. The second and more challenging attack is based on a live video attack where a genuine visual stream of the target person is coupled with a different audio recording of the target person. For this scenario we obtained an EER of 2.7%.

Comparing our results to previous work is not straightforward, since the type of data, definition of synchronization task, length of clips, and amount of training data differ. In general, most previous works obtain significantly higher EERs than ours, on longer clips.

The EERs reported in [11] are lower than ours; however, their goal to detect a temporal offset between audio and visual streams was different from ours.

The analysis of our results indicates a high correlation between synchrony detection errors and face recognition errors. Therefore, the impact of our synchrony detection errors is probably lower than presented.

Our ongoing work includes investigation of different scoring techniques, improvement of the deep learning framework, and further testing on other datasets.

6. References

[1] H. Aronowitz, M. Li, O. Toledo-Ronen, S. Harary, A. Geva, S. Ben-David, A. Rendel, R. Hoory, N. Ratha, S. Pankanti, and Others, "Multi-modal biometrics for mobile authentication," *Biometrics (IJCB), 2014 IEEE International Joint Conference on*, pp. 1—-8, 2014. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6996269

[2] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," *Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 8, 2014. [Online]. Available: http://www.cs.tau.ac.il/ wolf/papers/deepface_11_01_2013.pdf

[3] Y. Lei, N. Scheffer, L. Ferrer, and M. 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. [Online]. Available: http://www.sri.com/work/publications/novel-scheme-speaker-recognition-using-phonetically-aware-deep-neural-network

[4] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-End Text-Dependent Speaker Verification," no. Section 3, pp. 3–7, 2015. [Online]. Available: http://arxiv.org/abs/1509.08062

[5] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[6] Z. Wu, T. Kinnunen, N. Evans, and J. Yamagishi, "ASVspoof 2015: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan," pp. 1–5, 2014.

[7] E. Mordini, "TABULA RASA Trusted Biometrics under Spoofing Attacks," 2010. [Online]. Available: https://www.tabularasa-euproject.org/

[8] H. Aronowitz, R. Hoory, J. Pelecanos, and D. Nahamoo, "New developments in voice biometrics for user authentication," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2011.

[9] G. Chetty and M. 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.

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

[11] E. Marcheret, G. Potamianos, J. Vopicka, and V. Goel, "Detecting Audio-Visual Synchrony Using Deep Neural Networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[12] J. Hershey and J. Movellan, "Audio-vision: Using audio-visual synchrony to locate sounds," *Neural Information Processing Systems (NIPS'99)*, pp. 813–819, 1999.

[13] M. Slaney and M. Covell, "FaceSync: A linear operator for measuring synchronization of video facial images and audio tracks Malcolm," pp. 1–6.

[14] H. J. Nock, G. Iyengar, and C. Neti, "Assessing face and speech consistency for monologue detection in video," *Proceedings of the tenth ACM international conference on Multimedia - MULTIMEDIA '02*, p. 303, 2002.

[15] M. Gurban and J. P. Thiran, "Multimodal speaker localization in a probabilistic framework," *European Signal Processing Conference*, 2006.

[16] N. Eveno and L. Besacier, "A speaker independent" liveness" test for audio-visual biometrics." *Interspeech*, pp. 3081–3084, 2005.

[17] J. W. Fisher and T. Darrell, "Speaker association with signal-level audiovisual fusion," *IEEE Transactions on Multimedia*, vol. 6, no. 3, pp. 406–413, 2004.

[18] K. Kumar, J. Navrátil, E. Marcheret, V. Libal, and G. Potamianos, "Robust audio-visual speech synchrony detection by generalized bimodal linear prediction," in {*INTERSPEECH*} *2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009*. ISCA, 2009, pp. 2251–2254.

[19] R. Cutler and L. Davis, "Look who's talking: speaker detection using video and audio\ncorrelation," *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No.00TH8532)*, vol. 3, no. c, pp. 1589–1592, 2000.

[20] C. Bregler and Y. Konig, ""Eigenlips" for robust speech recognition," *Proceedings of ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. ii, pp. II/669–II/672, 1994.

[21] G. Feldhoffer, T. Bárdi, G. Takács, and A. Tihanyi, "Temporal asymmetry in relations of acoustic and visual features of speech," in *Signal Processing Conference, 2007 15th European*. IEEE, 2007, pp. 2341–2345.

[22] A. Karpov, A. Ronzhin, I. Kipyatkova, and M. Železnỳ, "Influene of phone-viseme temporal correlations on audiovisual stt and tts performance," 2011.

[23] H. Sakoe and S. 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.

[24] 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.

[25] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 511—-518, 2001.

[26] S. Milborrow and F. Nicolls, "Locating facial features with an extended active shape model," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5305 LNCS, no. PART 4, pp. 504–513, 2008.

[27] V. Kazemi and S. Josephine, "One Millisecond Face Alignment with an Ensemble of Regression Trees," *Computer Vision and Pattern Recognition (CVPR), 2014*, 2014. [Online]. Available: http://www.diva-portal.org/smash/record.jsf?pid=diva2:713097

[28] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 886–893, 2005.

[29] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a Similarity Metric Discriminatively, with Application to Face Verification," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 539–546.