VISUAL VOICE ACTIVITY DETECTION BASED ON SPATIOTEMPORAL INFORMATION AND BAG OF WORDS

Foteini Patrona, Alexandros Iosifidis, Anastasios Tefas, Nikolaos Nikolaidis, and Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
{tefas,nikolaid,pitas}@aiia.csd.auth.gr

ABSTRACT

A novel method for Visual Voice Activity Detection (V-VAD) that exploits local shape and motion information appearing at spatiotemporal locations of interest for facial region video description and the Bag of Words (BoW) model for facial region video representation is proposed in this paper. Facial region video classification is subsequently performed based on Single-hidden Layer Feedforward Neural (SLFN) network trained by applying the recently proposed kernel Extreme Learning Machine (kELM) algorithm on training facial videos depicting talking and non-talking persons. Experimental results on two publicly available V-VAD data sets, denote the effectiveness of the proposed method, since better generalization performance in unseen users is achieved, compared to recently proposed state-of-the-art methods.

***Index Terms*—** Voice Activity Detection, Space-Time Interest Points, Bag of Words model, kernel Extreme Learning Machine

1. INTRODUCTION

The task of identifying silent (vocal inactive) and non-silent (vocal active) periods in speech, called voice activity detection (VAD) has been widely studied for many decades using audio signals. In the last two decades, though, considerable attention has been paid to the use of visual information as an aid to the traditional Audio-only Voice Activity Detection (A-VAD), due to the fact that, contrary to audio, visual information is insensitive to environmental noise and can, thus, be of help to A-VAD methods for speech enhancement and speech source separation in noisy conditions.

The approaches proposed in the literature can be roughly divided in model-based and model-free ones, with the former requiring a training process, where positive and negative paradigms are employed for model learning and the latter not performing direct training, thus circumventing the need for an a-priori knowledge of the classes at the decision stage. Moreover, either visual or audiovisual data features can be exploited. In the latter case, combination of the two modalities can be achieved in two different ways, either by combining the features themselves (feature/early fusion) or by performing two separate uni-modal recognition stages and fusing their results (decision/late fusion).

Model-free V-VAD methods, usually rely solely on combinations of speaker-specific static and dynamic visual data parameters, like lip contour geometry and motion [1], or inner lip height and width trajectories [2] that are compared to appropriate thresholds for decision making. Emphasis is given on dynamic parameters due to the fact that identical lip shapes can be encountered both in silent and non-silent frames, making static features untrustworthy. In both these approaches, there is no discrimination between speech and non-speech acoustic events, which are thus handled as non-silent sections. Another model-free approach is proposed in [3], where signal detection algorithms are applied on mouth region pixel intensities along with their variations.

Concerning model-based V-VADs, features like lip opening, rounding and labio-dental touch (a binary feature indicating whether the lower lip is touching the upper teeth) for lip configuration followed by motion detection and SVM classification are proposed in [4], in an attempt to distinguish between moving and non-moving lips and then between lip motion originating either from speech or from other face/mouth activities, e.g., from facial expressions [1, 2]. Such a VAD system can constitute the first stage of a Visual Speech Recognition (VSR) system. The discriminative power of static and dynamic visual features in V-VAD is investigated in [5] where the predominance of dynamic ones is highlighted. The same approach is also adopted in [6], where facial profile as well as frontal views are used. Though not providing as much useful information as the frontal ones, facial profile views are proven to be useful in VAD.

An early-fusion model-based AV-VAD approach is introduced in [7]. 2D discrete cosine transformations (2D-DCTs) are extracted from the visual signal and a pair of GMMs is used for classification of the feature vector. V-VAD accuracy is quite high in the speaker-dependent case. However, it dramatically decreases in the speaker-independent case experiments, conducted on a simplistic dataset called GRID [8].

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 287674 (3DTVS). The European Union is not liable for any use that may be made of the information contained therein.

A new approach to V-VAD is introduced in this paper, regarding it as an action recognition problem. The Space Time Interest Point (STIP) detector [9] is utilized in order to detect video frame interest points that undergo abrupt intensity changes along space and time directions, Histogram of Oriented Gradient (HOG) and Histogram of Optical Flow (HOF) descriptors are calculated on each STIP video location and concatenated so that the Bag of Words (BoWs) model can be exploited in order produce a compact video representation. Classification is performed by applying a kernel Extreme Learning Machine classifier (kELM) [10, 11] for Single-hidden Layer Feedforward Neural (SLFN) network training. The proposed approach is evaluated on two publicly available data sets, namely GRID [8] and CUAVE [12] and experimental results denote that it can outperform recently proposed state-of-the-art methods on these databases.

The remainder of this paper is organized as follows. The proposed V-VAD approach is described in Section 2, the data sets used for evaluating our method performance and the respective experimental results are presented in Section 3 and conclusions are drawn in Section 4.

2. PROPOSED METHOD

The proposed method operates on grayscale facial region videos, extracted by applying face detection and tracking [13, 14] techniques. After determining the facial Regions of Interest (ROIs) in each video sequence, the facial ROIS are cropped and the resulting facial images are sized to fixed size of $H \times W$, thus producing the facial videos. The proposed V-VAD method is subsequently applied. In this Section, the proposed V-VAD method steps are described in detail, starting from the STIP-based facial video description.

2.1. STIP-based video description

Let $\mathcal{U}$ be an annotated facial video database containing N videos depicting human faces. In this paper, the Harris3D detector [15], a spatiotemporal extension of the Harris detector [16], is employed in order to detect video locations, where the image intensity values undergo significant spatiotemporal changes. After STIP localization, each facial video is described in terms of local shape and motion by a set of HOG/HOF descriptors (concatenation of l_2 normalized HOG and HOF descriptors), namely HOG/HOF feature vectors $\mathbf{p}_{ij} \in \mathbb{R}^D$, $i = 1,\ldots,N$, $j = 1,\ldots,N_i$, where i refers to the facial video index and j indicates the STIP index detected in facial video i. In all our experiments, the implementation [17] that is publicly available has been used and the dimensionality of the obtained HOG/HOF vectors is equal to $D = 162$.

2.2. BoW-based video representation

In the training phase, HOG/HOF vectors calculated on all training facial videos are employed in order to produce HOG/HOF vector prototypes, forming a codebook. This is achieved by applying K-Means clustering, so as to minimize the within-cluster scatter:

$$\sum_{i=1}^{N}\sum_{j=1}^{N_i}\sum_{k=1}^{K} \alpha_{ijk}\|\mathbf{p}_{ij} - \mathbf{z}_k\|^2, \tag{1}$$

where $\alpha_{ijk} = 1$, if $\mathbf{p}_{ij}$ is assigned to cluster k (having cardinality $n_k = \sum \alpha_{ijk}$) and $\alpha_{ijk} = 0$ otherwise. The codebook vectors $\mathbf{z}_k \in \mathbb{R}^D$, $k = 1,\ldots,K$ are determined to be the mean cluster vectors:

$$\mathbf{z}_k = \frac{1}{n_k}\sum_{i=1}^{N}\sum_{j=1}^{N_i} \alpha_{ijk}\mathbf{p}_{ij}. \tag{2}$$

The optimal action codebook cardinality (size) K is determined by applying a line search strategy, as will be explained in Section 3.

After codebook calculation, each action video can be represented by exploiting the similarity of the corresponding HOG/HOF vectors $\mathbf{p}_{ij}$ to each of the action codebook vectors $\mathbf{z}_k$. This is usually performed by applying hard vector quantization, i.e., by assigning each HOG/HOF vector $\mathbf{p}_{ij}$ to the closest action codebook vector $\mathbf{z}_k$ and thus determining the vectors $\mathbf{q}_i$, $i = 1,\ldots,N$, which are l_1 normalized in order to produce the facial vectors $\mathbf{s}_i$, i.e.,:

$$s_{ik} = \frac{q_{ik}}{\sum_{n=1}^{K} q_{in}}. \tag{3}$$

After the training facial vectors $\mathbf{s}_i$ have been obtained, they are employed for training a SLFN network by applying the kELM algorithm [10, 11], as will be described in the following.

2.3. Neural Network-based classification

After the facial vectors $\mathbf{s}_i$, $i = 1,\ldots,N$ have been calculated, they are employed, along with the corresponding talking/non-talking labels c_i, in order to train a SLFN network, which should consist of K input, L hidden and one output neurons, since the problem at hand is a two-class problem. The number L of hidden layer neurons is, usually, much greater than the number of classes involved in the classification problem [10, 18].

The network targets t_i, $i = 1,\ldots,N$, each corresponding to a facial vector $\mathbf{s}_i$, are set to $t_i = 1$ or $t_i = -1$, depending on whether the facial vector corresponds to a talking or a non-talking human face, respectively. In ELM-based classification schemes, the network input weights $\mathbf{W}_{in} \in \mathbb{R}^{K \times L}$ and the hidden layer bias values $\mathbf{b} \in \mathbb{R}^L$ are randomly assigned, while the network output weights $\mathbf{w} \in \mathbb{R}^{1 \times L}$ are analytically calculated. Let $\mathbf{v}_j$ and $\mathbf{w}_j$ denote the j-th column of $\mathbf{W}_{in}$ and the j-th element of $\mathbf{w}$, respectively. For a given activation function $\Phi()$, the output o_i of the SLFN network corresponding to the training action vector $\mathbf{s}_i$ is calculated by:

$$o_i = \sum_{j=1}^{L} w_j \, \Phi(\mathbf{v}_j, b_j, \mathbf{s}_i). \tag{4}$$

Many activation functions $\Phi()$ can be used for the calculation of the network hidden layer outputs, such as sigmoid, sine, Gaussian, hard-limiting and Radial Basis Functions (RBF). In our experiments, the $RBF - \chi^2$ activation function has been employed, as will be explained in Section 3.

By storing the network hidden layer outputs corresponding to the training facial vectors $\mathbf{s}_i$, $i = 1, \ldots, N$ in a matrix $\mathbf{\Phi}$,equation (4) can be expressed in a matrix form as $\mathbf{o} = \mathbf{\Phi}^T \mathbf{w}$ and by allowing small training errors in order to increase robustness to noisy data, the network output weight $\mathbf{w}$ can be obtained by solving:

$$\textbf{Minimize:} \quad \mathcal{J} = \frac{1}{2}\|\mathbf{w}\|_2^2 + \frac{c}{2}\sum_{i=1}^{N} \|\xi_i\|_2^2 \tag{5}$$

$$\textbf{Subject to:} \quad \mathbf{w}^T \boldsymbol{\phi}_i = t_i - \xi_i, \ i = 1, \ldots, N, \tag{6}$$

where ξ_i is the error corresponding to training facial vector $\mathbf{s}_i$, $\boldsymbol{\phi}_i$ is the i-th column of $\mathbf{\Phi}$ denoting the $\mathbf{s}_i$ representation in the ELM space and c is a parameter denoting the importance of the training error in the optimization problem. The optimal value of parameter c is determined by applying a line search strategy using cross-validation. The network output weight $\mathbf{w}$ is finally obtained by:

$$\mathbf{w} = \mathbf{\Phi}\left(\mathbf{K} + \frac{1}{c}\mathbf{I}\right)^{-1} \mathbf{t}, \tag{7}$$

where $\mathbf{K} \in \mathbb{R}^{N \times N}$ is the *ELM kernel matrix*, having elements equal to $[\mathbf{K}]_{i,j} = \boldsymbol{\phi}_i^T \boldsymbol{\phi}_j$ [11, 19].

By using (7), the network response o_l for a test vector $\mathbf{x}_l \in \mathbb{R}^D$ is given by:

$$o_l = \mathbf{W}_{out}^T \boldsymbol{\phi}_l = \mathbf{T}\left(\mathbf{\Phi}^T\mathbf{\Phi} + \frac{1}{c}\mathbf{I}\right)^{-1} \mathbf{k}_l, \tag{8}$$

where $\mathbf{k}_l \in \mathbb{R}^N$ is a vector having elements equal to $\mathbf{k}_{l,i} = \boldsymbol{\phi}_i^T \boldsymbol{\phi}_l$.

In most applications where ELM-based classification is performed, classification decision is made solely based on the sign of o_t. However, due to the fact that we are mainly interested in getting high precision values, i.e., high true positive rate, a threshold α was introduced in the training phase and fine tuning was performed in order to identify the threshold value giving the best precision values.

2.4. Facial video classification (test phase)

In the test phase, a test facial video is introduced to the proposed method. STIP video locations are detected and HOG/HOF descriptors are calculated, l_2 normalized and concatenated in order to form the corresponding HOG/HOF feature vectors $\mathbf{p}_{tj} \in \mathbb{R}^D$, $j = 1, \ldots, N_t$. $\mathbf{p}_{tj}$ are quantized by using the codebook vectors $\mathbf{z}_k \in \mathbb{R}^D$, $k = 1, \ldots, K$ produced in the training phase in order to determine the vector $\mathbf{q}_t \in \mathbb{R}^K$, which is l_1 normalized in order to produce the facial vector $\mathbf{s}_t$. $\mathbf{s}_t$ is subsequently introduced to the trained kELM network and its response o_t is obtained, resulting in the test facial video classification to the talking class if $o_t \geq \alpha$, or to the non-talking class if $o_t < \alpha$.

3. EXPERIMENTS

Experiments conducted in order to evaluate the performance of the proposed approach on visual voice activity detection are presented here. Two publicly available data sets, namely CUAVE and GRID, were used to this end, a short description of which is provided in the following subsections. Experimental results are subsequently given.

The optimal values of the parameters used in our method have been determined by applying a grid search strategy. That is, multiple experiments were conducted by using the values $c = 10^r$, $r = -6, \ldots, 6$ and $\alpha = 0.1e$, $e = 0, \ldots, 5$ and the best obtained performance is reported. The adoption of the training facial vectors $\mathbf{s}_i$, $i = 1, \ldots, N$ for the determination of the network hidden layer weights has been observed to provide satisfactory performance and, thus, $L = N$ was set and the training vectors were used for $\mathbf{W}_{in}$ in all the reported experiments. Due to the fact that the $RBF - \chi^2$ similarity metric provides the state-of-the-art performance [17, 18] for BoW-based video representations, $RBF - \chi^2$ activation function is used for the network hidden layer outputs calculation, i.e.:

$$\Phi(\mathbf{s}_i, \mathbf{v}_j, b) = exp\left(\frac{1}{2b}\sum_{d=1}^{D} \frac{(s_{id} - v_{jd})^2}{s_{id} + v_{jd}}\right), \tag{9}$$

where the value b is set equal to the mean χ^2 between the training data $\mathbf{x}_i$ and the network input weights $\mathbf{v}_j$.

3.1. The CUAVE Dataset

CUAVE [12] is a speaker-independent data set being used for voice activity detection, lip reading and speaker identification. It consists of videos of 36 male and female speakers (one video per speaker) with different skin complexions, accents and facial attributes, recorded both individually and in pairs uttering isolated and connected digits standing still or slightly moving in front of a simplistic background of solid color. The facial videos were extracted from the originals at a resolution of 195 × 315 pixels.

Experiments on this data set are usually conducted by performing multiple training-test rounds (sub-experiments), omitting a small percentage of the speakers and using 80% of the remaining for training and the rest 20% for testing, as suggested in [5, 6]. The performance of the evaluated method is subsequently measured by reporting the mean classification rate over all sub-experiments.

3.2. The GRID Dataset

The GRID corpus [8] is a collection of 34 male and female speakers uttering 1000 sentences each, standing perfectly still in front of a solid-color background. The sentences constituting the corpus are short and simple ones, of a standard syntax never encountered in real speech. The highest available data set resolution was selected for our experiments and the facial videos were extracted at a resolution of 300×300 pixels.

Both the speaker-dependent and speaker-independent experimental settings are widely adopted on this data set, as in [7], where the speaker-dependent experiment is conducted employing the videos of two speakers for training and those of another one for testing, while the speaker-independent using 80% of a speaker videos for training and the other 20% for testing.

3.3. Experimental Results

The proposed method has been applied on the two data sets by using the experimental protocols suggested in [5, 6, 7] and briefly described in subsections 3.1 and 3.2. It should be noted, though, that due to the fact that video-based classification is normally conducted by the proposed method, a preprocessing step of the data sets was necessary so that frame-based results were obtained and compared to those reported in the above papers. In addition, the facial video representation and description techniques used are such that when no or slight movements are encountered in a facial video, points of interest cannot be detected, and thus no descriptors are calculated either. Such facial videos are omitted from the classification process, but are taken into consideration in the calculations of the performance metrics used for method evaluation, considered as correctly classified visually silent examples.

More specifically, a sliding window of length equal to 7 frames moving with step equal to 1 frame was applied on the original videos of the CUAVE data set, splitting them in smaller parts and labels were assigned to the resulting videos using majority voting on the labels of the frames constituting them. Frame based classification, as in [5, 6], was thus performed, as the estimated labels were considered to refer to the middle frame of each short video. The sliding window length was chosen in such a way that the number of frames taking part in the classification of a video was equal to the number of frames used for the calculation of the dynamic features exploited by methods [5, 6]. A similar approach was also adopted for the GRID data set, where window of length 3 frames was used.

Comparison results with state-of-the-art methods evaluating their performance on the CUAVE and GRID data sets are provided in Tables 1 and 2, respectively. The metrics reported, are half total error rate (HTER) for the CUAVE data set and classification accuracy (CA) for the GRID data set. The performance of our previous method [3] is also provided. As can be seen, the proposed method outperforms the methods in [5, 6] obtaining half total error rates lower than half of those reported by them on the two experimental setups used in the CUAVE data set. On the other hand, method [3] performs surprisingly poorly, maybe due to the fact that the facial videos used were of a quite low quality, thus hindering its estimations, which are based on intensity values.

Table 1. Comparison results on the CUAVE dataset.

CUAVE (HTER)	Experiment [5]	Experiment [6]
Method [3]	47.1%	47.2%
Method [5]	25.6%	-
Method [6]	-	25.9%
Proposed method	**11.3%**	**11.7%**

As far as the GRID data set is concerned, the proposed method outperforms the method in [7] by 15.5% in the speaker-independent setup, highlighting its generalization ability in unseen users, while it achieves worse performance on the speaker-dependent experimental setting. The latter fact indicates the weakness of our method to exploit speaker-related features in order to enhance its performance, contrary to what is the case with method [7], whose performance is much better in the speaker-dependent experiment than in the speaker-independent. Moreover, method [3] also seems to perform quite better in this data set than in the CUAVE, achieving, in the speaker-independent experiment, classification accuracy only 4.2% lower than that reported in [7], which is not remarkably increased in the speaker-dependent setup, though.

Table 2. Comparison results on the GRID dataset.

GRID (CA)	speaker-independent experiment	speaker-dependent experiment
Method [3]	67.8%	68.3%
Method [7]	72.0%	**97.0%**
Proposed method	**87.5%**	87.4%

Overall, it can be seen that the proposed method achieves great generalization ability on new users, since in such experimental settings it outperforms the relating state-of-the-art methods [5, 7] in a large extend. However, it cannot efficiently exploit speaker-dependent features in order to achieve better classification results.

4. CONCLUSION

A novel method for Visual Voice Activity Detection exploiting local shape and motion information appearing at spatiotemporal locations of interest for facial video description and the BoW model for facial video representation was proposed in this paper. Neural Network-based classification based on the ELM classifier using the BoW-based facial video representations leads to satisfactory classification performance and experimental results on two publicly available data sets denote the effectiveness of the proposed method, since it outperforms recently proposed state-of-the-art methods in user independent experimental settings.

5. REFERENCES

[1] D. Sodoyer, B. Rivet, L. Girin, J-L. Schwartz, and C. Jutten, "An analysis of visual speech information applied to voice activity detection," *International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. I–I, 2006.

[2] D. Sodoyer, B. Rivet, L. Girin, C. Savariaux, J-L. Schwartz, and C. Jutten, "A study of lip movements during spontaneous dialog and its application to voice activity detection," *The Journal of the Acoustical Society of America*, vol. 125, no. 2, pp. 1184–1196, 2009.

[3] S. Siatras, N. Nikolaidis, and I. Pitas, "Visual speech detection using mouth region intensities," *European Signal Processing Conference*, 2006.

[4] K. Saenko, K. Livescu, M. Siracusa, K. Wilson, J. Glass, and T. Darrell, "Visual speech recognition with loosely synchronized feature streams," *International Conference on Computer Vision*, vol. 2, pp. 1424–1431, 2005.

[5] R. Navarathna, D. Dean, P. Lucey, S. Sridharan, and C. Fookes, "Dynamic visual features for visual-speech activity detection," *Conference of International Speech Communication Association*, 2010.

[6] R. Navarathna, D. Dean, S. Sridharan, C. Fookes, and P. Lucey, "Visual voice activity detection using frontal versus profile views," *International Conference on Digital Image Computing Techniques and Applications*, pp. 134–139, 2011.

[7] I. Almajai and B. Milner, "Using audio-visual features for robust voice activity detection in clean and noisy speech," *European Signal Processing Conference*, vol. 86, 2008.

[8] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421 – 2424, November 2006.

[9] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2–3, pp. 107–123, September 2005.

[10] G.B. Huang, Q.Y. Zhu, and C.K. Siew, "Extreme learning machine: a new learning scheme of feedforward neural networks," *International Joint Conference on Neural Networks*, vol. 2, pp. 985–990, July 2004.

[11] A. Iosifidis, A. Tefas, and I. Pitas, "On the kernel extreme learning machine classifier," *Pattern Recognition Letters*, 2014.

[12] E.K. Patterson, S. Gurbuz, Z. Tufekci, and J.N. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," *International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. II–2017 – II–2020, May 2002.

[13] G.N. Stamou, M. Krinidis, N. Nikolaidis, and I. Pitas, "A monocular system for person tracking: Implementation and testing," *Journal on Multimodal User Interfaces*, vol. 1, no. 2, pp. 31 – 47, 2007.

[14] O. Zoidi, A. Tefas, and I. Pitas, "Visual object tracking based on local steering kernels and color histograms," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 5, pp. 870 – 882, 2013.

[15] I. Laptev and T. Lindeberg, "Space-time interest points," *Internationa Conference on Computer Vision*, pp. 432–439, 2003.

[16] C. Harris and M. Stephens, "A combined corner and edge detector," *Alvey Vision Conference*, pp. 147–152, 1988.

[17] H. Wang, M.M. Ullah, A. Kläserr, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," *British Machine Vision Conference*, 2009.

[18] A. Iosifidis, A. Tefas, and I. Pitas, "Minimum class variance extreme learning machine for human action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 11, pp. 1968–1979, November 2013.

[19] G.B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 2, pp. 513–529, 2012.