IMPROVED AUDIO-VISUAL LAUGHTER DETECTION VIA MULTI-SCALE MULTI-RESOLUTION IMAGE TEXTURE FEATURES AND CLASSIFIER FUSION

Zahid Akhtar, Stefany Bedoya and Tiago H. Falk

INRS-EMT, University of Quebec, Canada

ABSTRACT

Efforts are afoot to design better context-aware human-computer interaction techniques that have knowledge of both their surrounding and the affective state of the user. One of the most important nonverbal behavioural cues for affective human-machine interaction is laughter. Automatic detection of laughter is an interesting, yet challenging problem, which in recent years has gained increased attention from both the academic and industrial communities. The majority of existing laughter detection systems rely on either audio or video modalities. Humans, however, typically rely on audio-visual cues during conversation and/or interaction, thus it is expected that improved results can be achieved if both modalities are used. In this work, we propose a multimodal framework that analyzes audio and video channels separately, then fuses their decisions. Conventional speech spectral and prosodic features are used, whereas new multi-scale multi-resolution binarized statistical image features are proposed due to their improved expressive power. Experiments with the publicly available MAHNOB Laughter database show that decision level fusion based on support vector machine classifiers leads to improved performance over single modality approaches, as well as over previously-proposed methods, all whilst requiring just a fraction of the computational power.

Index Terms— Laughter detection, information fusion, social signal processing, spectral features, BSIF features

1. INTRODUCTION

Social Signal Processing aims at bridging the social intelligence gap between humans and machines [1]. The last century has witnessed tremendous interest and progress in human behaviour understanding using different nonverbal social signals (e.g., facial expression, vocalizations, gesture, posture, etc.). Amongst all cues, laughter is one of the most vital nonverbal behavioural cues, since laughter is a prominent and very common signal in human communications. Automatic laughter recognition can be used to estimate the emotional/mental state of the person in affective computing [2, 4] applications, to identify non-speech segments in speech recognition systems, to analyze human-human nonverbal behaviour [1] or multi-party meeting [5], and for content-based multimedia tagging/retrieval [6].

The state-of-the-art in laughter detection is nascent. The majority of existing work has relied on audio information only [7], thus ignoring the visual information generated by facial expressions; something that humans use constantly in human-human interactions. Recently, some efforts have emerged to integrate audio and video modalities for laughter detection [9, 2, 6], but error rates are still fairly high [10]. Moreover, existing methods rely on complex features, exhibit high computational cost, and are very sensitive to train/test data mismatch (i.e., "in-the-wild" performance), thus hampering their use in everyday, real-time applications [11, 3].

In this paper, we aim to fill some of these limitations by exploring an audio-visual laughter detection algorithm based on spectral and prosodic audio features, as well as multi-scale, multi-resolution image features, namely the multi-scale binarized statistical image features (MBSIF) for video content extraction. To make a final multimodal prediction, the posterior probability scores obtained from support vector machine (SVM) classifiers, trained individually for audio and video channels, are fused using a decision level fusion method. Experimental results on the publicly available MAHNOB laughter database show that the proposed approach outperforms existing methods in audio-visual laughter detection.

The remainder of this paper is organized as follows. Section 2 presents prior works for laughter detection. The proposed approach is described in Section 3. Experimental protocol, dataset, and figures of merit are presented in Section 4. Experimental results and discussions are given in Section 5 and, finally, conclusions are drawn in Section 6.

2. RELATED WORK

Laughter is one of the key non-linguistic vocalizations and most frequently annotated nonverbal behaviour cues [5]. In recent years with the burgeoning of affective computing, interest in automatic laughter detection has increased. Predominantly, these studies have used only audio information for laughter recognition. For example, authors in [12] and [13] developed laughter detection methods based on spectral coefficients and phonetic features with hidden Markov models (HMM), respectively. Beke *et al.* [14] devised a multi-feature based laughter detector by constructing a corpus containing

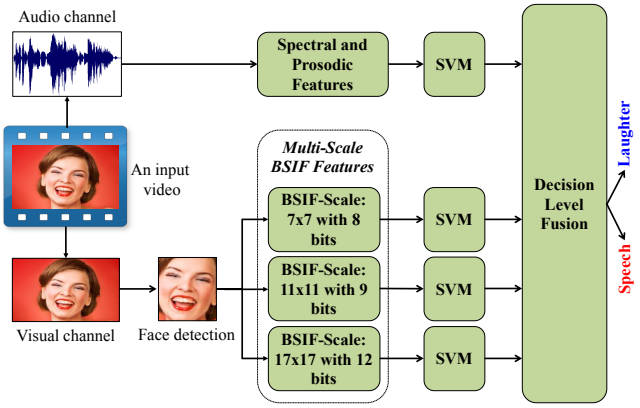

Fig. 1: Proposed multimodal laughter detection framework.

features such as mel frequency cepstral coefficients (MFCC), perceptual linear prediction (PLP), amongst others. In turn, Ringeval *et al.* [7] evaluated benchmark audio features from the 2010-2013 Interspeech Conference Challenges. All in all, spectral features outperformed prosodic ones [2, 7, 15]. Few works explored the use of face image/video information for laughter detection. For example, in [16] and [28], 113 and 10 facial points were explored and used as features, respectively. Only recently, has the fusion of audio-visual modalities for laughter detection been investigated. Existing works have relied on fusion of facial points as video features and speech spectral features [9, 8]. For instance, the method in [6] integrated 20 facial points visual features with six MFCC features and the speech signal zero-crossing rate (ZCR). More recently, [24] fused local binary pattern features with MFCC, pitch and jitter audio features, and in [29], a Kinect sensor was used to identify discriminative audio-visual features.

3. PROPOSED LAUGHTER DETECTOR

Laughter detection can be seen as a two-class classification problem, where the input video sample has to be flagged as either 'laughter' or 'speech'. The keynote of the process is to attain discriminant feature sets along with an appropriate classification scheme that accurately fuses decisions from audio and visual modalities. In this work, we propose a multimodal laughter detection method using multi-scale micro-texture video features, spectral and prosodic acoustic features and a decision-level fusion scheme, as depicted by Fig. 1.

The proposed framework first extracts information simultaneously from the audio and visual channels. In particular, the audio channel is used to extract spectral and prosodic features, respectively, from MFCCs and residual harmonics [23] based methods (described in Section 3.2). The visual features are extracted using binarized statistical image features (BSIF). In this work, we propose to use multiple BSIF filters with different scales in order to capture coarse texture, micro-texture and larger-scale visual laughter variation information for improved generalization. The multi-scale BSIF features are referred to as MBSIF and are described in Section 3.1.

As can be seen from the figure, extracted audio and visual features are then fed into modality-specific support vector machine (SVM) classifiers. The classification results of these individual SVMs are then combined by a decision level fusion scheme to obtain the final binary decision: laughter or speech. Specifically, the outputs of individual systems are compared, if they are not in agreement then the classifier with the highest likelihood score decides the final class prediction. More specifically, let $S_i(x)$ be the probability score given by ith classifier for sample x. The final decision can be estimated as $Decision = \text{argmax}[S_1(x), \cdots, S_k(x)]$, where $k = 4$ represents the number of classifiers explored herein.

3.1. Visual Features

Binarized statistical image features (BSIF) is a local image descriptor constructed by binarizing the responses to linear filters [17]. BSIF learns a set of filters from natural images using a ICA (independent component analysis) based unsupervised scheme. These learned filters are used to represent each pixel of the given image as a binary string by computing its response to learned filters. The binary string for each pixel can be considered as a local descriptor of the image intensity pattern in the neighbourhood of that pixel. Finally, the histogram of the pixels binary string values allows one to characterize the texture properties within the image sub-regions.

In this work, we have utilized the open-source filters [17], which were trained using 50000 image patches randomly sampled from 13 different natural scenic images [18]. Three main steps are needed to build the BSIF filters: mean subtraction of each patches, dimensionality reduction using PCA (principle component analysis), and estimation of statistically independent filters (or basis) using ICA. Given a visual laughter sample I of size $l \times m$ and a filter F_i of same size, the filter response is attained as follows [17]: $r_i = \sum_{l,m} I(l,m)F_i(l,m)$, where $F_i, \forall = \{1, 2, \ldots, m\}$ represents statistically independent filters whose response can be together calculated and binarized to obtain the binary string as [17]: $b_i = 1$ if $r_i > 0$, otherwise $b_i = 0$. Finally, the BSIF features are obtained as a normalized histogram of the pixel's binary codes, which can efficiently characterize the texture components in the laughter visual channel.

Filter size and length are important parameters to accurately identify visual laughter using BSIFs. Single filters with a fixed length may not be capable of generalizing well the visual laughters with varying intensities. Therefore, we propose to utilize multiple filters with different scales and resolutions in order to capture eminent features, thus the name multi-scale BSIF (MBSIF). In particular, we chose three different filters. The first is of size 17×17 with a length of 12 bits that will capture coarse texture information. The other two small scale filters (11×11 with a length of 9 bits and 7×7 with a length of 8 bits) will capture micro-texture information. These filters have been chosen based on overall performance of the

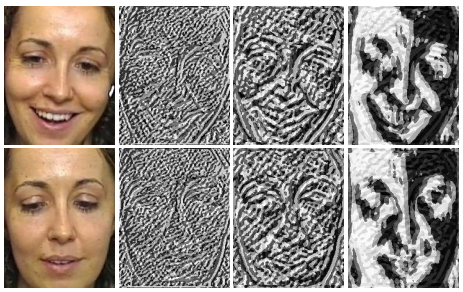

Fig. 2: Qualitative results of selected BSIF filters on laughter (top) and speech (bottom) samples from MAHNOB database. From left-to-right: Input image, BSIF features with 7×7, 11×11, and 17×17 filters.

proposed laughter detection system.

Figure 2 depicts the qualitative results of the three different filters used herein. As can be seen, features encoded by the 7×7 filter can be regarded as micro-textons, which are able to capture features especially for periorbital lines (i.e., laugh lines/crow's feet) and tear troughs. The 11×11 filter in turn, encapsulates macro-features, distinctly about nasolabial folds, oral commissures, and superficial lines. Lastly, the 17×17 filter captures bristly structures such as vermilion border, vertical lip lines and glabellar lines and so on. It is expected that by combining these different filters, thus capturing distinctive information that is important for laughter characterization, enhanced laughter detection will be achieved.

3.2. Audio Features

MFCCs have been widely used in audio-based laughter detection systems and various speech processing applications [16, 14, 2, 19]. MFCCs are compact representations of the speech signal and its spectral envelope [20]. In [21], it was shown that utilizing only 6 MFCCs could lead to similar laughter recognition performance as when using 12 coefficients, thus this parametrization has become popular within the field. Therefore, the same 6 coefficients are used here and were computed using 40 ms Hamming windows and 10 ms overlap. In addition to MFCCs, we also use pitch and jitter features to represent prosodic information. Pitch and jitter have been applied in several speech-based emotion recognition and speech discrimination tasks [22]. Some studies have claimed that for laughter detection spectral features are better than prosodic ones [2, 19]. However, higher pitch values are commonly observed during laughter [8] and recent work has shown advantages of combining spectral and prosodic features for better generalization capability [24]. As such, a similar approach is used here. More specifically, we utilize the pitch estimation scheme described in [23], which has been shown to be robust to noisy conditions. Pitch is measured in the range of 80-600 Hz using a frame length of 100 ms and frame shift of 10 ms. Jitter is then computed as the cycle-to-cycle variation of estimated pitch values. All audio features are extracted from the 16 KHz downsampled data.

4. EXPERIMENTAL SETUP

Here, we provide the experimental setup of the proposed audio-visual multimodal laughter detection framework.

4.1. Dataset

The MAHNOB Laughter Database [6] was used in this work, which comprises audio-visual information from 22 subjects (12 males and10 females). There are 563, 849, 51, 50 and 167 instances of spontaneous laughter, spontaneous speech utterances, acted laughter, posed smiles and vocalizations apart from speech and laughter, respectively. Videos were acquired at 25 frames per second using a camera with 720×576 pixels. Videos were compressed by the H.364 codec. Audio samples were recorded using built-in microphone in the camera with 2 channels, 48 KHz, 16 bits and a low signal-to-noise ratio.

4.2. Experimental Protocol and Figures-of-Merit

The final audio feature set used for classification was a 10-dimensional vector comprising the average 6-dimensional MFCC, and the average and standard deviation of the pitch and jitter parameters. For the visual channel, we used the Viola and Jones algorithm [25] to detect the face region for each frame. The following steps were then performed on the resulting facial region: conversion to grayscale, histogram equalization, and rescaling to 241×241. The SVM classifier with a radial basis function (RBF) kernel was used.

In this work, we followed same experimental protocol as in [6]. More specifically, experiments were conducted on a total of 554 laughter instances and 845 speech utterances. All experiments reported herein followed a leave-one-subject-out cross validation methodology, thus guaranteeing that the obtained results are subject independent. Results obtained in each fold are then averaged to get final results. Classification is performed on a frame-level basis by applying the classifier to all frames of a given episode resulting in a series of scores that are then combined into a final label.

The performance was evaluated via the F1-measure and overall classification rate (CR) as in [6, 16] and computed as: $F1 = (2 * precision * recall)/(precision + recall)$, $CR = (TP + TN)/(TP + TN + FP + FN)$, where TP, TN, FP and FN are true positives, true negatives, false positives and false negatives, respectively, and precision = (TP)/(TP + FP), and recall = (TP)/(TP + FN). Computational processing time for extracting visual features per sample/face was also used as a figure of merit. Processing time, as computed on Matlab version R2015a running on a desktop computer with an Intel Core i7-3770 processor, 3.40 GHz CPU and 16 GB of RAM, is used.

5. EXPERIMENTAL RESULTS AND DISCUSSION

The results of audio-only, video-only, and audio-visual laughter detection schemes are presented in Table 1 in terms of F1-

Algorithm	Modality	F1 (Laughter)	F1 (Speech)	CR
Petridis *et al.* [6]	Audio	77.2 ± 0.7	87.5 ± 0.3	83.9 ± 0.4
	Video	79.3 ± 1.1	89.7 ± 0.4	86.2 ± 0.6
	Audio-visual	86.5 ± 0.6	92.2 ± 0.3	90.1 ± 0.4
Bedoya & Falk [24]	Audio	90.1 ± 0.11	88.9 ± 0.07	91.9 ± 0.05
	Video	88.7 ± 0.15	71.7 ± 0.26	89.6 ± 0.08
	Audio-visual	92.7 ± 0.07	91.6 ± 0.06	93.3 ± 0.06
Proposed	Audio	90.1 ± 0.11	88.9 ± 0.07	91.9 ± 0.05
	Video†	84.4 ± 0.16	78.9 ± 0.19	83.4 ± 0.15
	Video‡	85.2 ± 0.11	77.0 ± 0.21	86.1 ± 0.10
	Video$^{\clubsuit}$	84.7 ± 0.13	75.9 ± 0.19	85.5 ± 0.10
	Video$^{\diamondsuit}$	86.3 ± 0.10	81.7 ± 0.17	87.2 ± 0.10
	Video$^{\heartsuit}$	87.8 ± 0.13	83.4 ± 0.15	88.3 ± 0.09
	Video$^{\spadesuit}$	86.5 ± 0.12	82.2 ± 0.17	88.3 ± 0.10
	Video$^{\oplus}$	86.2 ± 0.21	76.9 ± 0.10	86.0 ± 0.11
	Video*	87.3 ± 0.15	84.3 ± 0.17	88.5 ± 0.01
	Audio-visual†	89.0 ± 0.15	88.9 ± 0.07	90.0 ± 0.10
	Audio-visual‡	87.8 ± 0.14	90.2 ± 0.08	91.3 ± 0.09
	Audio-visual$^{\clubsuit}$	90.7 ± 0.11	90.0 ± 0.07	91.8 ± 0.08
	Audio-visual$^{\diamondsuit}$	87.3 ± 0.17	88.9 ± 0.08	89.9 ± 0.11
	Audio-visual$^{\heartsuit}$	88.9 ± 0.15	88.3 ± 0.08	89.9 ± 0.10
	Audio-visual$^{\spadesuit}$	87.2 ± 0.14	89.4 ± 0.08	90.8 ± 0.09
	Audio-visual$^{\oplus}$	91.4 ± 0.11	89.9 ± 0.07	91.8 ± 0.08
	Audio-visual*	94.5 ± 0.10	92.1 ± 0.08	94.3 ± 0.06

Table 1: Mean $\pm$ standard deviation of F1-measure and classification rates (CR) for laughter-vs-speech discrimination. Superscripts correspond to video features extracted using: †only 7×7 filter, ‡only 11×11 filter, $^{\clubsuit}$only 17×17 filter, $^{\diamondsuit}7 \times 7$ and 11×11 filters, $^{\heartsuit}7 \times 7$ and 17×17 filters, $^{\spadesuit}$11×11 and 17×17 filters, $^{\oplus}$all three filters with feature concatenation, *all three filters and decision level fusion, as in Fig. 1.

measure (%) and classification rates (CR-%). As can be seen, for video-only classification, the individual BSIF features are shown to achieve results inline with those previously reported in the literature. The gains arise when multiple filters are combined, thus incorporating multi-scale, multi-resolution details into the classifiers. Performances increase by combining the filters two-by-two and further gains are seen when all three filters are combined. Overall, the best performance is achieved with decision-level fusion where an 88.5% CR was achieved, thus outperforming the system proposed [6, 26] which exploits facial landmarks points features. Relative to the F1 (speech) score, the proposed fusion system outperformed the video-only method described in [24] (which relies on local binary pattern visual features) by 17.57%.

For the audio-only systems herein and in [24], results outperformed those achieved with video-only features and resulted in an overall CR of 91.9%. Such findings are expected as the audio channel conveys most discriminatory information and video conveys complementary information [2, 6]. Overall, the audio-only scheme used herein outperformed the video modality by 2.4% and obtained an improvement of 8.50% and 9.53% over the results reported in [26] and [6], respectively, thereby demonstrating the benefits of grouping spectral and prosodic features for the task at hand.

Audio-visual fusion, in turn, showed to result in the best overall performance. As with the video-only case, decision level fusion of the three BSIF features achieved the highest F1

	Predicted laughter	Predicted speech
Actual laughter	55.8 %	3.0 %
Actual speech	2.7 %	38.5 %

Table 2: Confusion matrix of proposed system

and CR scores of all compared methods. Relative to the individual video-only and audio-only accuracies, improvements of 6.57% and 2.62% could be seen, respectively, with the multimodal method. Such findings suggest that the proposed MBSIF features are indeed useful and robust for laughter detection, thus highlighting the importance of using multiple scale filters to capture prominent micro-texture (with small scale size) and coarse texture (using large scale size) information for laughter and speech discrimination.

Moreover, since MBSIF filters are designed using a set of natural image patches and ICA, it eradicates the need for manual tuning of filter parameters and maximizes the statistical independence between the learned filters ensuring effective information encoding. Moreover, the use of pre-learned filters removes the need for dataset/application specific learning. Moreover, since the proposed video features do not require facial points detection and tracking, lower computational complexity is required. For the experiments herein, computational complexity was inline with that achieved in [24] when all three filters were fused, but roughly 60% faster when using only one of the three filters. Relative to [6], the fused system was 54% faster and when using individual filters, 85% faster computational processing time was achieved.

Lastly, in Table 2 we report the percentage confusion matrix for the proposed multimodal approach. As can be seen, 3% of actual laughters are predicted as speech and 2.7% of actual speech is predicted as laughter. Closer inspection of these mistakes suggested that errors occurred when speech was very expressive (thus system confused as laughter) and when laughter episodes were created with mouths almost closed (system confused as speech). In such cases, the video stream conveyed contradictory information to what was expected by the classifier. Such issues may be corrected once larger datasets are made available, thus allowing the classifier to better learn how to handle such rare episodes.

6. CONCLUSION

This paper presented a novel multimodal framework for distinguishing laughter from speech using audio and visual information. More specifically, we proposed to use of multi-scale binarized statistical image features (MBSIF) fused with spectral and prosodic speech information. Experimental results show that the proposed multi-scale, multi-resolution features can outperform two other benchmarks using video-only features and achieve the highest performance overall when fused with speech information, all whilst requiring just a fraction of the computational power of the benchmark methods.

7. REFERENCES

[1] A. Vinciarelli, M. Pantic, H. Bourlard, Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, vol. 27, no. 12, pp. 1743-1759, 2009.

[2] S. Petridis, M. Pantic, Audiovisual discrimination between speech and laughter: Why and when visual information might help. *IEEE Transactions on Multimedia*, vol. 13, pp. 2, pp. 216–234, 2011.

[3] E. Sariyanidi, H. Gunes, and A. Cavallaro, Automatic analysis of facial affect: A survey of registration, representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[4] S. Zhalehpour, O. Onder, Z. Akhtar, C. Eroglu Erdem, BAUM-1: A Spontaneous Audio-Visual Face Database of Affective and Mental States. *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 300-313, 2017.

[5] S. Cosentino, S. Sessa, A. Takanishi, Quantitative laughter detection, measurement, and classification: A critical survey. *IEEE Reviews in Biomedical Engineering*, 2016.

[6] S. Petridis, B. Martinez, M. Pantic, The MAHNOB Laughter database. *Image and Vision Computing*, vol. 31, no. 2, pp. 186-202, 2013.

[7] T. Jacykiewicz *et al.*, Automatic Recognition of Laughter Using Verbal and Non-Verbal Acoustic., *Master's thesis, Universit de Fribourg*, 2014.

[8] S. Petridis, M. Pantic, J. F. Cohn, Prediction-based classification for audiovisual discrimination between laughter and speech. *IEEE FG Workshops*, pp. 619–626, 2011.

[9] S. Escalera, E. Puertas, P. Radeva and O. Pujol, Multimodal laughter recognition in video conversations. *IEEE CVPR Workshops*, pp. 110-115, 2009.

[10] H. J. Griffin et al., Perception and Automatic Recognition of Laughter from Whole-Body Motion: Continuous and Categorical Perspectives. *IEEE Trans. on Affective Computing*, vol. 6, no. 2, pp. 165-178, 2015.

[11] H. Salamin, A. Polychroniou, A. Vinciarelli, Automatic Detection of Laughter and Fillers in Spontaneous Mobile Phone Conversations. *IEEE Int'l Conf. on Systems, Man, and Cybernetics*, pp. 4282–4287, 2013.

[12] A. Lockerd, F. Mueller, LAFCam: Leveraging Affective Feedback Camcorder. *In CHI*, pp. 574-575, 2002.

[13] N. Campbell, H. Kashioka, R. Ohara, No laughing matter. *In European Conf. on SCT*, pp. 465-468, 2005.

[14] T. Neuberger, A. Beke, Automatic laughter detection in spontaneous speech using GMM-SVM method. *Int'l Conf. on Text, Speech, and Dialogue*, pp. 113–120, 2013.

[15] M. Graciarena, E. Shriberg, A. Stolcke, F. Enos, J. Hirschberg and S. Kajarekar, Combining Prosodic Lexical and Cepstral Systems for Deceptive Speech Detection. *IEEE ICASSP*, 2006.

[16] S. Petridis, M. Leveque, and M. Pantic, Audiovisual detection of laughter in human-machine interaction, *In Conf. on Affective Comp. and Int. Int.*, pp. 129–134, 2013.

[17] J. Kannala and E. Rahtu, BSIF: Binarized statistical image features. *In ICPR*, pp. 1363-1366, 2012.

[18] A. Hyvearinen, J. Hurri, and P. O. Hoyer, *Natural Image Statistics*, vol. 39, Springer-Verlag, 2009.

[19] K. P. Truong and D. A. Van Leeuwen, Automatic discrimination between laughter and speech, *Speech Communication*, (49)2:144–158, 2007.

[20] S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Tran. on Acoustics, Speech, and Sig. Proc.*, vol. 28, no. 4, pp. 357-366, 1980.

[21] L. S. Kennedy and D. P. Ellis, Laughter detection in meetings. *NIST ICASSP Workshop*. pp. 118–121, 2004.

[22] M. El Ayadi, M. S. Kamel, F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[23] T. Drugman and A. Alwan, Joint robust voicing detection and pitch estimation based on residual harmonics. *Interspeech*, pp. 1973–1976, 2011.

[24] S. Bedoya, T. Falk, Laughter Detection based on the Fusion of Local Binary Patterns, Spectral and Prosodic Features. *IEEE Int'l Works. on Multimedia Sig. Pro.*, 2016.

[25] P. Viola and M. Jones, Rapid object detection using a boosted cascade of simple features. *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 511-518, 2001.

[26] O. Rudovic, S. Petridis, M. Pantic, Bimodal Log-linear Regression for Fusion of Audio and Visual Features. *ACM Int'l Conf. on Multimedia*, pp. 789–792, 2013.

[27] S. Petridis and M. Pantic, Audiovisual laughter detection based on temporal features. *Int'l conf. on Multimodal interfaces*, pp. 37-44, 2008.

[28] A. Ito, X. Wang, M. Suzuki, and S. Makino, Smile and laughter recognition using speech processing and face recognition from conversation video. *International Conference on Cyberworlds*, pp. 8-14, 2005.

[29] B. B. Turker, Z. Buinca, M. T. Sezgin, Y. Yemez and E. Erzin, Real-time audiovisual laughter detection. *Signal Processing and Comm. Applications Conf.*, pp. 1-4, 2017.