Laughter Detection based on the Fusion of Local Binary Patterns, Spectral and Prosodic Features

Stefany Bedoya and Tiago H. Falk
INRS-EMT, University of Québec, Montréal, QC, Canada

***Abstract*—Today, great focus has been placed on context-aware human-machine interaction, where systems are aware not only of the surrounding environment, but also about the mental/affective state of the user. Such knowledge can allow for the interaction to become more human-like. To this end, automatic discrimination between laughter and speech has emerged as an interesting, yet challenging problem. Typically, audio- or video-based methods have been proposed in the literature; humans, however, are known to integrate both sensory modalities during conversation and/or interaction. As such, this paper explores the fusion of support vector machine classifiers trained on local binary pattern (LBP) video features, as well as speech spectral and prosodic features as a way of improving laughter detection performance. Experimental results on the publicly-available MAHNOB Laughter database show that the proposed audio-visual fusion scheme can achieve a laughter detection accuracy of 93.3%, thus outperforming systems trained on audio or visual features alone.**

***Index Terms*—speech, laughter, spectral features, cepstral features, local binary patterns**

I. Introduction

Laughter is a common non-verbal vocalization that also includes posture and facial activity changes around the mouth, cheeks, and often the upper face [1]. Laughter is usually described as a predominant and involuntary behaviour in social interaction. As such, automated laughter detection has emerged as a useful tool to improve human-machine interaction to make it more human-like, typically by sensing the user's affective state and adapting decisions accordingly [2], [3]. Moreover, laughter detection has also found its way into automatic speech recognizers, as a way of improving the performance for spontaneous speech [4]. Since laughter can be caused by different emotional states, it is a challenging task.

Typically, discrimination between speech and laughter is achieved with audio features alone. In [5], for example, a large pool of benchmark audio features from the 2010-2013 Interspeech Conference Challenges and features set based on formants were tested. In [6], in turn, features included 13-th order mel frequency cepstral coefficients (MFCCs) and perceptual linear prediction (PLP) coefficients, along with their delta (Δ) and double-delta ($\Delta\Delta$) coefficients, fundamental frequency (F0), jitter, shimmer, local root mean square energy (RMS), harmonic-to-noise-ratio (HNR), zero crossing rate (ZRC), spectral slope and LP center of gravity, amongst several others. Feature selection was performed based on receiver operating characteristic (ROC) curves and the area under curve (AUC) measure. Overall, spectral features have been shown to outperform prosodic ones and both on decision level and feature level improves the final performance [7], [8], [9], [10]. More recently, integration of audio-visual features for laughter detection has been explored where the majority of the published works have focused on facial points and speech spectral features. In [7], for example, MFCCs, Δ, energy, pitch and ZCR were combined with 20 facial point (rigid-face movement and nonrigid-face-movement points) features extracted using PatrasPantic particle filtering tracking scheme. Audio features outperformed video ones and the fusion of the two modalities improved performance by 2.6%. In [1], in turn, 113 facial point features were used together with 13 MFCC and Δ coefficients. Similarly, audio features outperformed video-based ones and fusion of the two resulted in improvements of 1.3%. In two follow-up studies [11], [12], 6 MFCCs (and ZRC in [8]) were used as audio features and 20 facial points for visual features. Under such setup, gains attained from multimodal fusion were as high as 4.5% and an overall accuracy of 90.1% was achieved for speech vs. laughter discrimination. While the majority of existing papers have focused on facial points to characterize visual information for laughter detection, recent work has suggested that local binary pattern (LBP) video features can be useful for facial expression recognition and emotion detection [13], [14], [15]. As such, this paper explores the use of LBP texture features extracted from mouth and eye regions, as these facial regions provide useful information for laughter vs speech discrimination. More specifically, the LBP texture features are extracted from each region independently and are then concatenated to better encode appearance and spatial relations of facial regions [13]. The proposed features are tolerant to illumination changes and are computationally simpler than the commonly-used facial point features. From audio, in turn, MFCCs and spectral features are used.

For the task of laughter detection/discrimination, previous unimodal work has explored the use of Gaussian mixture model (GMM), hidden Markov model (HMMs), and support vector machine (SVM) classifiers (see [16], [17], [18], [19] for more details) with the latter standing out as the top-performer. Moreover, for audio-visual detection, feature- and decision-level fusion schemes have been proposed with the latter typically resulting in improved performances [11], [12]. Following these insights, the present work proposes the fusion of the decisions from SVM classifiers trained on audio-alone and video-alone features using a competing scores scheme.

The remainder of paper is organized as follows. Section II presents audio-visual features used in our experiments. Section III describes the experimental setup, followed by the Results and Discussion in Section IV. Lastly conclusions are presented

in Section V.

II. Multimodal Laughter Detection: Features

This section presents the audio-visual features used in the proposed laughter detection algorithm.

A. Visual Features

Local Binary Patterns (LBP) is a model of texture analysis, where a texture image can be characterized by its texture spectrum [20]. The operator labels the examined window into cells, for each pixel in a cell, compares the pixel to each of 3×3 neighbourhoods and the results are considered as a binary number. The 256-bin histogram of the LBP computed over a region is used as a texture descriptor. To be able to deal with textures at different scales, the LBP operator was extended to use circular neighbourhoods [13], [21]. A circularly symmetric neighbour set is composed for a neighbourhood of P sampling points on a circle of radius R ($R > 0$) denoted $LBP_{P,R}$ (see Fig. 1). The P ($P > 1$) sampling points produce 2^P local binary patterns to describe a texture image. Invariant texture operator T in a local neighbourhood of a monochrome texture image can be described in function of gray levels of P, such as in [21]:

$$T = t(g_c, g_0, ..., g_{P-1}), \tag{1}$$

where gray value g_c corresponds to the gray value of center pixel of the local neighbourhood and $g_p(p = 0, ..., P = 1)$ corresponds to gray values of P.

An alternate extension of LBPs use uniform patterns [15]. An LBP is called uniform if the binary pattern contains at most two bitwise transitions from 0 to 1 or vice versa when the bit pattern is considered circular [13]. The final histogram has a separate bin for every uniform pattern and all non-uniform patterns are assigned to a single bin. When there are more than two transitions into a single bin, it is possible to accumulate the transitions in an operator denoted by $LBP^{u2}_{P,R}$, with less than 2^p bins. This way, the number of labels for a neighbourhood of 8 pixels can be reduced from 256 to 59 for LBP^{u2} [13]. Using the uniform patterns, we can account nearly 90% of all patterns in the (8,1) neighbourhoods [21].

In our work, we follow the feature extraction approach described in [13]. More specifically, the Viola and Jones algorithm was used to detect mouth and eye regions for each frame [22]. The resulting facial regions were adjusted to increase the contrast, then converted to grayscale and scaled to 59×97 and 60×89 pixels, respectively. Finally, 59 labels were calculated for each extracted region in an neighbourhood of 8 pixels using the $LBP^{u2}_{P,R}$ operator with uniform patterns. In total, 177 (3×59) descriptors were computed and concatenated into a single feature histogram to be used for classification. Figure 2 depicts an example of the original face image and the detected mouth and eye regions.

Fig. 2. Example of original face image and the cropped eye and mouth regions

B. Audio Features

As mentioned above, MFCCs have dominated audio-based laughter detection systems and have been successfully used across numerous speech applications [1], [6], [7], [8], [9]. MFCCs are compact representations of the speech signal and its spectral envelope [23]. In [24], it was shown that using only the first 6 MFCC, similar laughter detection perfomance could be achieved to systems based on 12 coefficients. As such, in our experiments, only 6 coefficients are used and were computed using 40 ms Hamming windows and 10 ms overlap.

In addition to MFCCs, pitch and jitter features were used to characterize prosodic information. Pitch and jitter have been used in speech-based emotion recognition and are commonly used for other speech discrimination tasks (e.g., [25], [26]). While previous work has suggested that spectral features are better than prosodic ones for laughter detection [7], [8], we have decided to use both feature types, as subjective testing has reported that higher pitch values are present during laughter than during speech [27]. The pitch estimation algorithm described in [28] was used, as it was shown to be more reliable in noisy conditions. Pitch was measured in the range of 80-600 Hz using a frame length of 100 ms and frame shift of 10ms. All audio features were extracted from 16kHz downsampled speech/laughter data. The final audio feature set used for classification was a 10-dimensional vector comprised of the average 6-dimensional MFCC, combined with the average and standard deviation of the pitch and jitter parameters.

III. Experimental setup

In this section, brief descriptions of the database and classification methodology are presented.

A. Database

In our experiments, the MAHNOB Laughter Database was used [11]. The database contains audio-visual information from 22 subjects (12 males, 10 females) totalling 563 instances of spontaneous laughter, 849 spontaneous speech utterances, 51 instances of acted laughter, approximately 50 instances of posed smiles, and 167 vocalizations other than speech and laughter. Video recordings were made at 25 frames per second

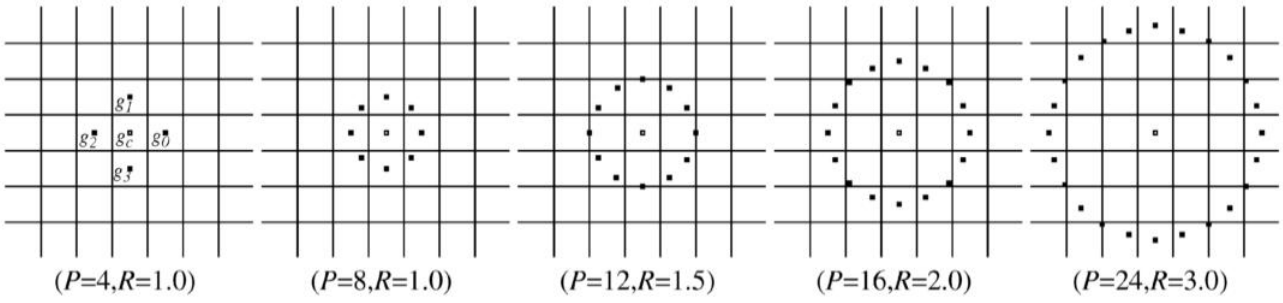

Fig. 1. Examples of the extended LBP for various (P,R) values (from [16])

Fig. 3. Sample images from MAHNOB Laughter dataset. Top corresponds to laughter segments while bottom images to speech utterances.

using a video camera which has a resolution of 720×576 pixels and samples were compressed using the H.364 codec. Voice samples were obtained using the built-in microphone in the camera (2 channels, 48 kHz, 16 bits) with a low signal-to-noise ratio.

The recordings labelled as laughter and speech were used for our experiment. A total of 13 instances (9 laughter, 4 speech) were excluded in concordance with [11]. As such, a total of 554 laughter instances and 845 speech utterances were processed. The list of annotations is administered by the authors of the MANHOB database. Figure 3 depicts sample images taken from the MAHNOB database during laughter (top) and speech (bottom).

B. Classifier Design and Figures-of-Merit

Laughter vs speech discrimination is performed using a support vector machine (SVM) classifier [29] with a radial basis function (RBF) kernel, as it has shown to provide the best performance over other kernels in previous emotion and laughter discrimination tasks [14], [30]. To estimate the generalization performance of the classifiers, we perform a leave-one-subject-out cross-validation methodology. For each fold, data from one subject is considered to be unseen and used as the testing set, while data from the remaining subjects is used for training.

As figures-of-merit, two parameters are used: the per class F1-measure and the overall classification rate (CR). Both are reported as percentages and are computed as:

$$F1 = \frac{2 * precision * recall}{precision + recall}, \tag{2}$$

$$CR = \frac{TP + TN}{TP + TN + FP + FN}, \tag{3}$$

where TP, TN, FP and FN are true positives, true negatives, false positives and false negatives, respectively, and

$$precision = \frac{TP}{TP + FP}, \tag{4}$$

$$recall = \frac{TP}{TP + FN}. \tag{5}$$

C. Fusion Scheme

Here, decision-level fusion is performed based on a competing scores scheme of the decisions made by the audio-only and video-only SVM classifiers. More specifically, the outputs of the individual systems are compared and if they are not in agreement, the classifier with the highest likelihood score decides the final class prediction. Thus, if $c_k(x_i)$ corresponds to the prediction score for sample x_i of each classifier k, then the final prediction score can be computed as:

$$C = \arg\max[c_1(x_i) \dots c_k(x_i)], \tag{6}$$

where $k = 2$ represents the number of classifiers in the fusion scheme.

IV. Experimental Results and Discussion

A. Results

Table I presents the per class F1-measure and classification rates (CR) for each experiment using the audio-only, video-only, and audio-visual schemes. As can be seen, the simple fusion scheme improved increase by 2.2% and 3.7% compared with audio-only and video-only rates, respectively. This increase is inline with those previously reported in the literature. The overall CR of 93.3%, however, was higher than those reported in the literature for the same dataset and experimental setup [11].

	F1 (Laughter)	F1 (Speech)	CR
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)

TABLE I
MEAN (AND STANDARD DEVIATION) OF THE F1 AND CLASSIFICATION RATES (CR) FOR LAUGHTER-VS-SPEECH DISCRIMINATION

	Predicted laughter	Predicted speech
Actual laughter	54.9%	0.8%
Actual speech	5.9%	38.4%

TABLE II
CONFUSION MATRIX FOR AUDIOVISUAL CLASSIFICATION

B. Discussion: Unimodal laughter detection

For video-only classification, we obtained a CR of 89.6% using the proposed LBP features. These results outperform those reported in [11] for the same dataset (and leave-one-subject-out experimental methodology), but which relied on facial point features instead. Such findings suggest that LBP features are indeed useful for laughter detection from videos and should be considered in future systems.

Mouth and eyes regions have been widely used in speech and emotion recognition due to geometric-type features variations, such as height, width and area [31], [32]. In our work, the LBP texture features were extracted using these areas. it allowed to minimize the number of LBP histograms used to represent facial expression proposed in previous studies [13], [14], where the face images were equally divided into small areas to extract the LBP coefficients.

Moreover, as in previous studies, our results show that audio-only classification outperformed video-based classification and achieved an overall CR of 91.9%. Since laughter can be accompanied by subtle facial expressions, this can explain the fact that visual information can be less representative for the discrimination task than the information contained in voice features [7]. Overall, the audio-only scheme outperformed the video one by 2.3%. Moreover, an improvement of 2.1% over the results reported in [33] show the benefits of grouping spectral and prosodic features for the task at hand.

C. Discussion: Multimodal laughter detection

Laughter and speech are both audio-visual events that contain representative information from each modality for the discrimination task. As such, it is expected that audio-visual fusion results in improved performance. This was indeed the case and the proposed decision-level fusion scheme outperformed those reported in [11] by 3.2%. In general, the visual scheme provides information which cannot be corrupted by acoustic noise in the environment and therefore may improve the final performance.

The confusion matrix for the audiovisual approach are shown in Table II . The main source of confusion for the final scheme is speech utterances being classifier as laughter episodes, whereas a few of laughter episodes being classifier as speech utterances. In our work, we believe that the main explanation for the better performance for laughter over speech events is the facial expression of the subject, a smile produced by subjects helps the audiovisual system. However, when the speech is produced with a smile-like expression, then the visual information does not help the system.

V. CONCLUSIONS

This paper has explored the use of local binary pattern (LBP) video features for laughter detection, as well as their fusion with spectral and prosodic speech information for multimodal laughter detection. Experimental results on a publicly-available dataset show the LBP features achieving superior performance relative to conventional facial point features. Once fused with audio features via a simple decision level fusion scheme, improvements of up to 3.7% could be achieved relative to using the video features alone. Overall, a classification rate of speech versus laughter detection of 93.3% was achieved, thus outperforming results previously reported in the literature for the same dataset and experimental setup.

ACKNOWLEDGEMENTS

The authors with to acknowledge funding from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

[1] S. Petridis, M. Leveque, and M. Pantic, "Audiovisual detection of laughter in human-machine interaction," in *Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, 2013, pp. 129–134.

[2] J. A. Russell, J. Bachorowski, and J. M. Fernndez-Dols, "Facial and vocal expressions of emotion," *Annual Review of Psychology*, vol. 54, no. 1, pp. 329–349, 2003.

[3] M. Pantic, A. Pentland, A. Nijholt, and T. S. Huang, "Human computing and machine understanding of human behavior: a survey," in *Artifical Intelligence for Human Computing*. Springer, 2007, pp. 47–71.

[4] A. Vinciarelli, N. Suditu, and M. Pantic, "Implicit human-centered tagging," in *IEEE International Conference on Multimedia and Expo*, 2009, pp. 1428–1431.

[5] T. Jacykiewicz and F. Ringeval, "Automatic recognition of laughter using verbal and non-verbal acoustic," Master's thesis, Universit de Fribourg.

[6] T. Neuberger and A. Beke, "Automatic laughter detection in spontaneous speech using gmm–svm method," in *Text, Speech, and Dialogue*. Springer, 2013, pp. 113–120.

[7] S. Petridis and M. Pantic, "Audiovisual discrimination between speech and laughter: Why and when visual information might help," *IEEE Transactions on Multimedia*, vol. 13, no. 2, pp. 216–234, 2011.

[8] K. P. Truong and D. A. Van Leeuwen, "Automatic discrimination between laughter and speech," *Speech Communication*, vol. 49, no. 2, pp. 144–158, 2007.

[9] M. Graciarena, E. Shriberg, A. Stolcke, F. Enos, J. Hirschberg, and S. Kajarekar, "Combining prosodic lexical and cepstral systems for deceptive speech detection," in *IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1, 2006, pp. I–I.

[10] B. Yin, E. Ambikairajah, and F. Chen, "Combining cepstral and prosodic features in language identification," in *18th IEEE International Conference on Pattern Recognition*, vol. 4, 2006, pp. 254–257.

[11] S. Petridis, B. Martinez, and M. Pantic, "The mahnob laughter database," *Image and Vision Computing*, vol. 31, pp. 186–202, 2013.

[12] S. Petridis, M. Pantic, and J. F. Cohn, "Prediction-based classification for audiovisual discrimination between laughter and speech," in *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*, 2011, pp. 619–626.

[13] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[14] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803 – 816, 2009.

[15] M. Kächele, M. Glodek, D. Zharkov, S. Meudt, and F. Schwenker, "Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression," *depression*, vol. 1, no. 1, 2014.

[16] B. Reuderink, M. Poel, K. Truong, R. Poppe, and M. Pantic, ch. Decision-Level Fusion for Audio-Visual Laughter Detection.

[17] E. Osuna, R. Freund, and F. Girosi, "Training support vector machines: an application to face detection," in *IEEE computer society conference on Computer vision and pattern recognition*, 1997, pp. 130–136.

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

[19] M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kächele, M. Schmidt, H. Neumann, G. Palm *et al.*, "Multiple classifier systems for the classification of audio-visual emotional states," in *Affective Computing and Intelligent Interaction*. Springer, 2011, pp. 359–368.

[20] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[21] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

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

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

[24] L. S. Kennedy and D. P. Ellis, "Laughter detection in meetings," in *NIST ICASSP 2004 Meeting Recognition Workshop*. National Institute of Standards and Technology, 2004, pp. 118–121.

[25] T. Vogt, E. André, and J. Wagner, "Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation," in *Affect and emotion in human-computer interaction*. Springer, 2008, pp. 75–91.

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

[27] S. Petridis, M. Pantic, and J. F. Cohn, "Prediction-based classification for audiovisual discrimination between laughter and speech," in *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*, 2011, pp. 619–626.

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

[29] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[30] A. Sayedelahl, R. Araujo, and M. S. Kamel, "Audio-visual feature-decision level fusion for spontaneous emotion estimation in speech conversations," in *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2013, pp. 1–6.

[31] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *Computer vision-eccv 2004*. Springer, 2004, pp. 469–481.

[32] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[33] M. Graciarena, E. Shriberg, A. Stolcke, F. Enos, J. Hirschberg, and S. Kajarekar, "Combining prosodic lexical and cepstral systems for deceptive speech detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 2006, pp. I–I.