Decision-Level Fusion for Audio-Visual Laughter Detection

Boris Reuderink[1], Mannes Poel[1], Khiet Truong[1,2], Ronald Poppe[1], and Maja Pantic[1,3]

[1] University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands
{reuderin,m.poel,truongkp,poppe,panticm}@ewi.utwente.nl
[2] TNO Defence, Sec. and Safety, P.O. Box 23, 3769 ZG Soesterberg, The Netherlands
[3] Imperial College, Dept. of Computing, 180 Queen's Gate, London SW7 2AZ, UK

Abstract. Laughter is a highly variable signal, which can be caused by a spectrum of emotions. This makes the automatic detection of laughter a challenging, but interesting task. We perform automatic laughter detection using audio-visual data from the AMI Meeting Corpus. Audio-visual laughter detection is performed by fusing the results of separate audio and video classifiers on the decision level. This results in laughter detection with a significantly higher AUC-ROC[1] than single-modality classification.

1 Introduction

Laughter is omnipresent in human vocal communication, and conveys cues for emotional states. This makes automatic laughter detection an interesting research subject. Earlier work on laughter detection has mainly focused on laughter detection in audio. In this work, we will add the video modality, and perform audio-visual laughter detection. We will construct classifiers for the audio and video modalities independently, and test if fusion of these modalities can improve the performance of automatic laughter detection.

In the next section we will describe some previous research on laughter detection and fusion of audio-visual data. Then we will outline the experiment, present our results and end with conclusions and suggestions for future work.

2 Previous Work

2.1 Laughter Detection in Audio

Automatic laughter detection has been studied several times in the context of meetings, for audio indexing and to detect affective states. We will describe a number of studies on automatic laughter detection in audio, and summarize some characteristics of these studies.

[1] Area under curve - receiver operating characteristic.

A. Popescu-Belis and R. Stiefelhagen (Eds.): MLMI 2008, LNCS 5237, pp. 137–148, 2008.

Campbell et al. developed a system to classify a laugh in different categories [3]. They constructed a corpus from daily speech containing four affective classes of laughter: a hearty laugh, an amused laugh, a satirical laugh and a social laugh. A training set of 3000 hand-labeled laughs was used to train Hidden Markov Models (HMMs). The HMMs recognized the affective class correctly in 75% of the test cases. Automatic laughter detection is frequently studied in the context of meetings. Kennedy and Ellis [13] detected multiple laughing participants in the ICSI Meeting Corpus. Using a Support Vector Machine (SVM) on one second windows of Mel-Frequency Cepstrum Coefficients (MFCCs) features, an equal error rate (EER) of 13% was obtained. Truong and Van Leeuwen [21] used a clean subset of the ICSI Meeting Corpus to train Gaussian Mixture Model (GMM) and SVM classifiers. Instances containing speech and inaudible laughs were removed to form the clean subset. The classifiers were trained on spectral features, pitch & energy, pitch & voicing features and modulation-spectrum features. Usually, the SVM classifiers performed better than the GMM classifiers. Fusion based on the output of the GMM and SVM classifiers increases the discriminative power, as does fusion between classifiers based on spectral features and classifiers based on prosodic information.

When we compare the results of these studies, GMMs and SVMs seem to be used most for automatic laughter recognition. Spectral features seem to outperform prosodic features, and although different corpora are used, an EER of 12–13% seems to be usual.

2.2 Audio-Visual Fusion

Most work on audio-visual fusion has focused on the detection of emotions [2, 9, 10, 25, 27]. Some other studies perform cry detection [15], movie classification [24], tracking [1], speech recognition [6] and laughter detection [12]. These studies all try to exploit the complementary nature of audio-visual data. Decision-level fusion is usually performed using the product, or a (weighted) sum of the predictions of single-modality classifiers. As an alternative to decision-level fusion, sometimes feature-level fusion is used where the features are merged before classification. An overview of relevant work on audio-visual fusion can be found in Table 1.

Audio-visual laughter detection has already been performed by Ito et al. [12] on a database with Japanese, English and Chinese subjects. The lip lengths, the lip angles and the mean intensities of the cheek areas were used as features for the video modality. Frame level classification of the video features was performed using a perceptron, resulting in a recall of 71%, and a precision of 52%. Laughter sound detection was performed on MFCC and delta-MFCC features, using two GMMs, one for laughter, and one for other sounds. A recall of 96% and a precision of 60% was obtained using 16 Gaussian mixtures. Decision-level fusion was performed with manually designed rules, resulting in a recall of 71% and a precision of 74%. Ito et al. do not report if this increase is statistically significant.

Recently, Petridis and Pantic performed audio-visual discrimination between laughter and speech [17]. The AMI Meeting database was used to create a corpus

Table 1. Audio-visual fusion. The last column contains the performance using different modalities and fusion techniques; A indicates audio, V indicates video, FF indicates feature-level fusion, and DF indicates decision-level fusion. The performance is measured in classification accuracy, except for [12, 17, 18] for which we present the F_1 measure instead of recall - precision pairs.

Study	Dataset	Performance
Petridis and Pantic [17] (2008)	AMI, spontaneous, laughter	A: $F_1 = 0.64$, V: $F_1 = 0.80$, DF: $F_1 = 0.82$, FF: $F_1 = 0.81$
Petridis and Pantic [18] (2008)	AMI, spontaneous, laughter	A: $F_1 = 0.69$, V: $F_1 = 0.80$, DF: $F_1 = 0.88$
Zeng et al. [26] (2007)	AAI, spontanous, 2 emotions	A: 70%, V: 86% DF: 90%
Hoch et al. [10] (2005)	Posed, 3 emotions	A: 82%, V: 67%, DF: 87%
Ito et al. [12] (2005)	Spontaneous, laughter	A: $F_1 = 0.72$, V: $F_1 = 0.60$, DF: $F_1 = 0.72$
Wang and Guan [23] (2005)	Posed, 6 emotions	A: 66%, V: 49%, FF: 82%
Busso et al. [2] (2004)	Posed, 4 emotions	A: 71%, V: 85%, FF: 89%, DF: 89%
Go et al. [8] (2003)	Unknown, 6 emotions	A: 93, V: 93%, DF: 97%
Dupont and Luettin [6] (2000)	M2VTS, spontaneous, 10 words	A: 52% V: 60%, FF: 70%, MF: 80%, DF: 82%

with 40 laughter segments and 56 speech segments. These laughter segments contain a clearly audible harmonic laugh, and do not contain speech. Video features were extracted by tracking 20 facial points, and transformed to uncorrelated features using a PCA similar to our approach in [19]. A few relevant principal components were used to calculate distance based features. Perceptual Linear Prediction coding (PLP) was used to obtain audio-features. For classification, AdaBoost was used to select a feature-subset, on which an Artificial Neural Network classifier was trained. Both decision-level and feature-level fusion of the audio and video modality seem to improve on the performance of the video-classifier slightly (see Table 1) but it remains to be seen on which level fusion works best. In a follow-up study Petridis and Pantic use the same dataset to perform decision-level fusion based on different configurations of single-modality classifiers, such as spectral and pitch & energy based audio-classifiers, and face-component and head-component based video-classifiers [18]. The best combination was formed by the combination of the spectral audio-classifier and both the head and face modality for video.

From Table 1 it appears that fusion of the audio and video modality boosts the classification performance generally with a few percent. However, most work does not report the significance of this gain in performance. The fusion of audio and video modalities seems to work best when the individual modalities both have a low performance, for example due to noise in the audio-visual speech recognition of Dupont [6]. When single classifiers have a high performance, the

performance gain obtained by fusion of the modalities is low, and sometimes fusion even degrades the performance, as observed in the work of Gunes and Piccardi [9].

3 Methodology

We perform fusion on the decision-level where the audio and video modalities are classified separately. When the classifiers for both modalities have classified the instance, their results are used to make a final multi-modal prediction. We have chosen to evaluate decision-level fusion because it allows us to use different classifiers for each of the two modalities.

3.1 Dataset

Previous work on laughter detection often used the ICSI Meeting Corpus. Because this corpus does not provide video recordings, we have created a dataset based on the AMI Meeting Corpus. The AMI Meeting Corpus consists of 100 hours of meeting recordings, stored in different signals that are synchronized to a common time line. The meetings are recorded in English, mostly spoken by non-native speakers. For each meeting, there are multiple audio and video recordings. We used seven unscripted meetings recorded in the IDIAP-room (IB4001, IB4002, IB4003, IB4004, IB4005, IB4010, IB4011) as these meetings contain a fair amount of spontaneous laughter. We removed two of the twelve subjects; one displayed extremely asymmetrical facial expressions (IB4005.2), the other displayed a strong nervous tick in the muscles around the mouth (IB4003.3, IB4003.4). We used the close-up video recording (DivX AVI codec 5.2.1, 2300 Kbps, 720×576 pixels, 25 frames per second) and the headset audio recording (16 KHz WAV file) of each participant for our corpus.

We were unable to use the laughter-annotations provided with the AMI-Corpus as these are often not correctly aligned. Therefore the seven meetings we selected from the AMI Meeting Corpus were segmented into laughter by the first author. Due to the spontaneous nature of these meetings, speech, chewing and occlusions sometimes co-occur with the laughter and non-laughter segments.

The final corpus is built from the segmented data. The laughter instances are created by padding each laughter segment with 3 seconds on each side to capture the visual onset and offset of a laughter event. Laughter segments that overlapped after padding are merged into a single laughter instance. A preliminary experiment indicated that including these 3 seconds improved the classification performance significantly. The non-laughter instances are created from the audio-visual data that remains after removing all the laughter segments. The length of the non-laughter instances is taken from a random Gaussian distribution with a mean and standard deviation equal to the mean and standard deviation of the laughter segments.

We have based our corpus on 60 randomly selected laughter and 120 randomly selected non-laughter instances, in which 20 facial points needed for tracking are

Fig. 1. Example laughter segments for the subjects

visible. We included some barely audible laughs and laughter overlapping with speech, in contrast to [17, 18] where no speech was included in the laughter segments. Some examples of laughter segments are displayed in Fig. 1. We made sure no smiles occurred in the non-laughter instances. To test the validity of the class-labels, two other annotators annotated the corpus. One annotator rated 4 laughter-instances as non-laughter, the other annotator agreed completely, resulting in a agreement of 97.7%. Of all the 180 instances, 59% contains speech of the visible participant. Almost all instances contain background speech. Together these instances form 25 minutes of audio-visual data. The dataset is available at `http://hmi.ewi.utwente.nl/ami-laughter`.

3.2 Features

Audio Features. We use RASTA-PLP features to encode the audio-signal. RASTA-PLP adds filtering capabilities for channel distortions to PLP features, and yields significantly better results for speech recognition tasks in noisy

environments than PLP [6]. We used the same settings as were used by Truong and Van Leeuwen for PLP features [21]. The 13 cepstral coefficients (12 model order, 1 gain) are calculated over a window of 32 ms with a step-size of 16 ms. Combined with the temporal derivative (calculated by convolving with a simple linear-slope filter over 5 audio frames) this resulted in a 26 dimensional feature vector per audio frame. We normalized these 26-dimensional feature vectors to a mean $\mu = 0$ and a standard deviation $\sigma = 1$ using z-normalization.

Video Features. The video channel is transformed into sequences of 20 two-dimensional facial points located on key features of the human face. These point sequences are subsequently transformed into orthogonal features using a Principal Component Analysis (PCA).

The points are tracked as follows. The points are manually assigned at the first frame of an instance movie and tracked using a tracking scheme based on particle filtering with factorized likelihoods [16]. We track the brows (2 points each), the eyes (4 points each), the nose (3 points), the mouth (4 points) and chin (1 point). This results in a compact representation of the facial movement in a movie using 20 (x, y)-tuples per frame. This tracking configuration has been used successfully for the detection of the atomic action units of the Facial Action Coding System (FACS) [22].

After tracking, we performed a PCA on the 20 points per video-frame without reducing the number of dimensions; the principal components now serve as a parametric model, similar to the Active Shape Model of Cootes et al. [5]. No label information was used to create this model. An analysis of the eigenvectors revealed that the first five principal components encode the head pose, including translation, rotation and scale. The other components encode interpersonal differences, facial expressions and corrections for the linear approximations of movements (see Figure 3.5 of [19]).

In order to capture temporal aspects of this model, the first order derivative for each component is added to each frame. The derivative is calculated with $\Delta t = 4$ frames on a moving average of the principal components with a window length of 2 frames. Again, we normalized this 80-dimensional feature vector to a mean $\mu = 0$ and a standard deviation $\sigma = 1$ using z-normalization.

3.3 Classification

We evaluate Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs) and Support Vector Machines (SVMs) for classification. GMMs and HMMs model the distribution for both classes and classify by estimating the probability that an instance was produced by the model for a specific class. GMMs and HMMs are frequently used in speech recognition and speaker identification, and have been used before for laughter recognition [3, 12, 14, 21]. SVMs are discriminatory classifiers, and have been used for laughter detection in [13, 21]. We used HMMs and GMMs for the audio-modality and SVMs for the video-modality as this resulted in the best performance [19].

The HMMs we use model the generated output using a mixture of Gaussian distributions. We used two different topologies; the left-right HMMs that are

frequently used in speech recognition, and ergodic HMMs that allow transitions from all states to all states. For the SVMs we use a sliding window of 1.20 seconds to create fixed-length features from the video segments. During classification, a probability estimate for the different windows of an instance is calculated. The final prediction of an instance is the mean of its window-predictions. We use Radial Basis Function (RBF) kernel SVMs, which are trained using LIBSVM [4].

To estimate the generalization performance of the classifiers, we perform two times 15-fold cross-validation. Inside each fold, we use 1/28 of the training data as a validation set to select model parameters such as the HMM configuration, the number of Gaussians and the C and γ parameter of the SVMs, the rest of the training data is used to train classifiers. To find well-performing model parameters we use a multi-resolution grid search [11]. Note that we extracted the PCA-model outside of the cross-validation loop to focus on the generalization performance of the classification. However, we do not expect that this has a big influence on the measured performance.

Fusion. Fusion is performed on the decision-level, which means that the output of an audio and a video classifier is used as input for the final fused prediction. For each instance we classify, probability estimates are generated for the audio and video modalities. Fusion SVMs are trained on the z-scores of the estimates using the same training, validation and test sets as used for the single modality classifiers. The output of these SVMs is a multi-modal prediction based on high-level fusion. As an alternative to this learned fusion, we tested fusion using a weighted-sum of the single-modality predictions:

$$f_{\text{fused}}(x) = \alpha * f_{\text{video}}(x) + (1 - \alpha) * f_{\text{audio}}(x). \quad (1)$$

Evaluation. We have chosen to use the Area Under Curve of the Receiver Operating Characteristic (AUC-ROC) as performance measure because it does not depend on the bias of the classifier, and is class-skew invariant [7]. The AUC-ROC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. In addition to the AUC-ROC performance, we will report the EER for a classifier. The EER is the point on the ROC where the false-positive rate equals the false-negative rate. A paired two-tailed t-test is used to compare the AUC-ROCs of the different classifiers.

4 Results

For audio, the GMM classifiers performed better than the HMM classifiers, resulting in a mean AUC-ROC of 0.825. On average 16.9 Gaussian mixtures were used to model laughter, non-laughter was modeled using 35.6 Gaussian mixtures. The HMM performed slightly worse with an AUC-ROC of 0.822. The HMMs used 11.6 fully connected states to model laughter, and 21.3 fully connected states to model non-laughter. Surprisingly, no left-right HMMs were selected in

Table 2. The performance of the audio and video classifiers. The standard deviation of the AUC-ROCs is displayed between parenthesis.

Classifier	Params	AUC-ROC	EER
RASTA-GMM	16.9 (3.2) pos. mix., 35.6 (5.9) neg. mix.	0.825 (0.143)	0.258
RASTA-HMM	11.6 (1.9) pos. states, 21.3 (1.9) neg. states	0.822 (0.135)	0.242
Video-SVM	$C = 2.46$, $\gamma = 3.8 \times 10^{-6}$	0.916 (0.114)	0.133

Table 3. Results of the decision-level fusion. The t-test is a paired samples t-test on the AUC-ROCs of the video-SVM (V-SVM) classifier and the specified fusion classifiers. The mean value of the AUC-ROCs is displayed with the standard deviation displayed between parenthesis.

Fusion	Features	T-test	AUC-ROC	EER
RBF-SVM	V-SVM + R-GMM	$t(29) = 2.45, p < 0.05$	0.928 (0.107)	0.142
RBF-SVM	V-SVM + R-HMM	$t(29) = 1.93, p = 0.06$	0.928 (0.104)	0.142
W-sum, $\alpha = 0.57$	V-SVM + R-GMM	$t(29) = 2.69, p < 0.05$	0.928 (0.107)	0.142
W-sum, $\alpha = 0.55$	V-SVM + R-HMM	$t(29) = 2.38, p < 0.05$	0.930 (0.101)	0.142

the model selection procedure. This indicates that there was no strict sequential pattern for laughter that could be exploited for recognition, which seems to support the claim that laughter is a group of sounds [20].

The SVM video-classifier outperformed the audio-classifiers with an AUC-ROC of 0.916, using a mean $C = 2.46$ and a mean $\gamma = 3.8 \times 10^{-6}$. See Table 2 for the performance of the different single-modality classifiers. Note that these performances are measured on normalized datasets, and we do not test the generalization performance over subjects.

We used these classifiers to perform decision-level fusion. The performance of the different fusion configurations is displayed in Table 3. The fused classifiers have a higher mean AUC-ROC than the single-modality classifiers. In the case of SVM-fusion, the combination of the video-SVM classifier and the RASTA-GMM classifiers outperforms the best single-modality classifier slightly, but significantly. Inspection of the trained (RBF) SVM-classifiers reveals that the separating hyperplane is nearly linear.

In addition to fusion using a SVM, we used a weighted-sum rule (1) to combine the output of the audio and video classifiers. The weight of both modalities is determined using the α parameter. The highest mean AUC-ROC values are obtained in the region with a more dominant audio-classifier. However, for a significant improvement over the video-SVM classifier $\alpha = 0.57$ and $\alpha = 0.55$ are needed for the RASTA-GMM and the RASTA-HMM classifier respectively (see Table 3).

When we compare the ROC of the linear fusion classifiers with the ROC of the video-SVM classifiers, we can see that the EER of the fused classifiers is higher than the EER of the video-SVM classifiers (see Fig. 2). Most of the performance-gain is obtained in the direct vicinity of the EER point, where the error-rates are not equal. This trend is also visible with the SVM-fusion. This can

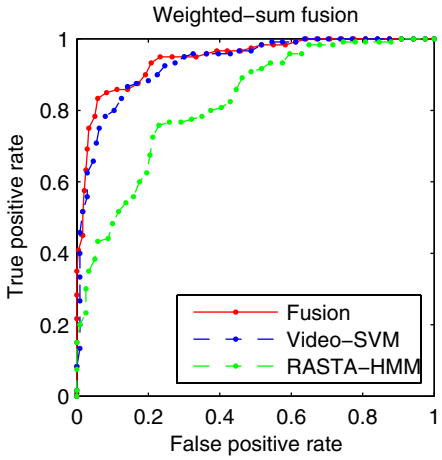

Fig. 2. ROCs for the video-SVM, RASTA-HMM and weighted sum ($\alpha = 0.55$) fusion

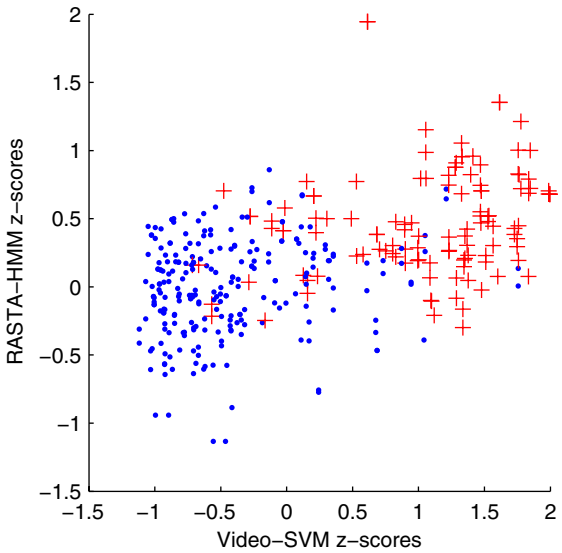

Fig. 3. The normalized output of the audio and video classifiers on the test-sets. Laughter instances are marked with an plus, non-laughter instances are marked with a dot.

be explained by the observation that for unequal error rates the fusion classifier can exploit the complementary nature of both modalities, which it cannot do for the threshold with an equal error rate, where the hyperplane needs to separate instances for which both modalities are uncertain (see Fig. 3).

5 Conclusion and Future Work

Our goal was to perform automatic laughter detection by fusing audio and video signals on the decision level. We have built audio and video-classifiers, and demonstrated that the fused classifiers significantly outperformed the best single-modality classifiers. The best audio-visual classifiers are constructed using a weighted sum of the RASTA-HMM and video-SVM classifiers, resulting in a AUC-ROC performance of 0.930. While fusion on the decision-level improves the performance of the laughter-classifier significantly, fusion seems only beneficial for classification with unequal false-negative and false-positive rates. With equal error rates, the decision-boundary has to separate instances for which both modalities are uncertain. For unequal error rates, these instances fall on one side of the decision-boundary, and now instances with only one uncertain modality can be classified more reliably, resulting in a better performance.

For future work we recommend an investigation of fusion on the feature-level. We have demonstrated that decision-level fusion can improve the performance, but it is not yet clear how this relates to other fusion techniques, such as feature-level fusion. Previous work on audio-visual laughter detection is inconclusive on this subject. A limitation of this experiment is that we removed smiles from our corpus. Adding a smile class to the corpus would most likely decrease the performance of the video-classifier. A follow-up experiment could show if fusion would increase the performance in this setting. In addition to these technical challenges, focussing on the context in which laughter and smiles occur would form an interesting subject. During segmentation we observed interaction between laughter and smiles of different participants in a meeting. It is likely that laughter detection can be improved by explicit use of interactions and semantic information.

References

[1] Beal, M.J., Jojic, N., Attias, H.: A graphical model for audiovisual object tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 25(7), 828–836 (2003)

[2] Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S.: Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Proceedings of the International Conference on Multimodal Interfaces (ICME 2004), State College, PA, October 2004, pp. 205–211 (2004)

[3] Campbell, N., Kashioka, H., Ohara, R.: No laughing matter. In: Proceedings of the Interspeech, Lisbon, Portugal, September 2005, pp. 465–468 (2005)

[4] Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001), `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

[5] Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models - their training and application. Computer Vision and Image Understanding (CVIU) 61(1), 38–59 (1995)

[6] Dupont, S., Luettin, J.: Audio-visual speech modeling for continuous speech recognition. IEEE Transactions on Multimedia 2(3), 141–151 (2000)

[7] Fawcett, T.: An introduction to ROC analysis. Pattern Recognition Letters 27(8), 861–874 (2006)
[8] Go, H.-J., Kwak, K.-C., Lee, D.-J., Chun, M.-G.: Emotion recognition from the facial image and speech signal. In: Proceedings of the SICE Annual Conference, Fukui, Japan, August 2003, vol. 3, pp. 2890–2895 (2003)
[9] Gunes, H., Piccardi, M.: Fusing face and body display for bi-modal emotion recognition: Single frame analysis and multi-frame post integration. In: Tao, J., Tan, T., Picard, R.W. (eds.) ACII 2005. LNCS, vol. 3784, pp. 102–111. Springer, Heidelberg (2005)
[10] Hoch, S., Althoff, F., McGlaun, G., Rigoll, G.: Bimodal fusion of emotional data in an automotive environment. In: Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP 2005), Philadelphia, PA, vol. 2, pp. 1085–1088 (2005)
[11] Hsu, C.-W., Chang, C.-C., Lin, C.-J.: A practical guide to support vector classification. Technical report, National Taiwan University, Taipei, Taiwan (July 2003)
[12] Ito, A., Wang, X., Suzuki, M., Makino, S.: Smile and laughter recognition using speech processing and face recognition from conversation video. In: Proceedings of the International Conference on Cyberworlds (CW 2005), Singapore, November 2005, pp. 437–444 (2005)
[13] Kennedy, L.S., Ellis, D.P.W.: Laughter detection in meetings. In: Proceedings of the NIST Meeting Recognition Workshop at the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP 2004), Montreal, Canada (May 2004)
[14] Lockerd, A., Mueller, F.L.: Leveraging affective feedback camcorder. In: Extended abstracts of the Conference on Human Factors in Computing Systems (CHI 2002), Minneapolis, MN, April 2002, pp. 574–575 (2002)
[15] Pal, P., Iyer, A.N., Yantorno, R.E.: Emotion detection from infant facial expressions and cries. In: Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP 2006), Toulouse, France, May 2006, vol. 2, pp. 721–724 (2006)
[16] Patras, I., Pantic, M.: Particle filtering with factorized likelihoods for tracking facial features. In: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG 2004), Seoul, Korea, pp. 97–102 (2004)
[17] Petridis, S., Pantic, M.: Audiovisual discrimination between laughter and speech. In: Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP 2008), Las Vegas, NV, pp. 5117–5120 (2008)
[18] Petridis, S., Pantic, M.: Fusion of audio and visual cues for laughter detection. In: Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR 2008), Niagara Falls, Canada (to appear, 2008)
[19] Reuderink, B.: Fusion for audio-visual laughter detection. Technical report, University of Twente (2007)
[20] Trouvain, J.: Segmenting phonetic units in laughter. In: Proceedings of the International Conference of the Phonetic Sciences, Barcelona, Spain, August 2003, pp. 2793–2796 (2003)
[21] Truong, K.P., van Leeuwen, D.A.: Automatic discrimination between laughter and speech. Speech Communication 49(2), 144–158 (2007)
[22] Valstar, M.F., Pantic, M., Ambadar, Z., Cohn, J.F.: Spontaneous vs. posed facial behavior: automatic analysis of brow actions. In: Proceedings of the International Conference on Multimodal Interfaces (ICME 2006), Banff, Canada, November 2006, pp. 162–170 (2006)

[23] Wang, Y., Guan, L.: Recognizing human emotion from audiovisual information. In: Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP 2005), Philadelphia, PA, vol. 2, pp. 1125–1128 (2005)
[24] Xu, M., Chia, L.-T., Jin, J.S.: Affective content analysis in comedy and horror videos by audio emotional event detection. In: Proceedings of the International Conference on Multimodal Interfaces (ICME 2005), Amsterdam, The Netherlands, July 2005, pp. 622–625 (2005)
[25] Zajdel, W., Krijnders, J., Andringa, T., Gavrila, D.: CASSANDRA: Audio-video sensor fusion for aggression detection. In: Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS 2007), London, United Kingdom, September 2007, pp. 200–205 (2007)
[26] Zeng, Z., Hu, Y., Roisman, G.I., Wen, Z., Fu, Y., Huang, T.S.: Audio-visual spontaneous emotion recognition. Artifical Intelligence for Human Computing, 72–90 (2007)
[27] Zeng, Z., Tu, J., Liu, M., Huang, T.S., Pianfetti, B., Roth, D., Levinson, S.: Audio-visual affect recognition. IEEE Transactions on Multimedia 9(2), 424–428 (2007)