Noise Analysis in Audio-Visual Emotion Recognition

Ntombikayise Banda
University of Cambridge
15 JJ Thompson Avenue
Cambridge, CB3 0FD
nb395@cam.ac.uk

Peter Robinson
University of Cambridge
15 JJ Thompson Avenue
Cambridge, CB3 0FD
pr10@cam.ac.uk

ABSTRACT

This paper describes the use of a decision-based fusion framework to infer emotion from audiovisual feeds, and investigates the effect of noise on the fusion system. Facial expression features are constructed from linear binary patterns, and are processed independently of the prosodic features. A linear support vector machine is used for the fusion of the two channels. The results show that the recognition accuracy of the bimodal system improves on the individual channels; moreover, the system maintains a reasonably good performance in the presence of noise.

1. INTRODUCTION

Recent studies indicate that there is a fundamental interdependence between emotion and cognition, that is, emotion can influence both the *process* of thinking (*how* we deal with information) and the *content* of thinking and behaviour (*what* we think and do) [2]. Therefore, with human-computer interaction (HCI) researchers seeking to make systems efficient, more user-friendly and receptive to users' needs, it follows that emotion can no longer be excluded from the research.

It has been shown that emotion can be inferred from facial expressions, body posture, speech utterances and physiological signs, for example, the pulse, heart rate and skin conductance [7]. The motive for multimodal research can be found in how the human brain recognizes emotion. It has been found that the brain bases its decision on integrated information emanating from multiple sensors. According to Campanella [1] there is clear empirical evidence that the brain integrates information from face and voice. Campanella further states that it is advantageous for the brain to integrate these two sources of information for the following reasons: (i) by exploiting redundancies between the face and voice, it increases the reliability of sensory estimates, (ii) by combining non-redundant, complementary cues it maximizes information gathered from the two modalities.

The work presented in this paper is based on the same database used by Haq et al. [5] who investigated the fusion of the audio and video channels at feature and decision levels. The faces of the actors used in the evaluation were painted with markers (see Figure 1 for sample image). The authors extracted the facial features by manually labelling the first frame of each video sequence, and then using a marker tracker to track the x and y positions of the painted markers. These positions together with audio features were passed through a feature selection algorithm, after which linear transformation techniques were applied for feature reduction. The audio and video features were combined using Gaussian classifiers at different fusion levels. The algorithm yielded a 98% emotion classification rate based on the audio-visual features, 53% for audio features and 98% for visual features.

We adopt a decision-based fusion approach which integrates decision outputs from the facial expression and vocal affect analysis subsystems. The facial expression analysis subsystem encodes each image in the video sequence as linear binary patterns from which feature vectors are constructed. Classification is achieved through Random Forests. The audio analysis system extracts prosodic features such as the pitch and speech rate to characterize emotions, and classifies them using pairwise support vector machines. Classifier fusion is employed for the integration of the two channels. This work is discussed in section 3.2.1.

We extend the study to investigate the effect of noise on the emotion recognition system. Although most of the audio-visual recordings used are made under quiet, laboratory conditions, the target applications will typically be deployed in environments with cluttered backgrounds and various levels of ambient noise. It is therefore important to study noise effects to determine if noise-reduction mechanisms should be considered in the design of emotion recognition systems. However, very little research has been conducted in this area. The few works that analyze noise effects in emotion recognition have been limited to the speech signal. Schuller et al [8] investigated emotion estimation from noisy speech (additive white noise and noise resulting from different audio capturing techniques) and showed how different acoustic feature sets adapt to the noisy conditions. You et al [10] proposed an enhanced Lipschitz embedding algorithm for emotion analysis and classification, and compared it with other dimensionality reduction methods. Their results show how the proposed method consistently outperfomed other methods

even when the speech signal was infused with white noise at various signal-to-noise ratio (SNR) levels. We differ in that we analyze the effect of a corrupt video signal on affect estimation. Our observations are summarized in section 3.2.2.

2. METHODOLOGY

2.1 Database

The analysis in this paper is based on the Surrey Audio-Visual Expressed Emotion (SAVEE) database [5]. The database consists of four actors of ages 27 to 31 depicting the widely-used six basic emotions (anger, disgust, fear, happiness, sadness and surprise), plus the neutral state. The recordings consist of 15 phonetically-balanced TIMIT sentences per emotion (with 15 additional sentences for neutral state) resulting in a corpus of 480 British English utterances. The actors' frontal face are painted with 60 markers as means for facial feature extraction for the work presented in [5]. This study will however not make use of the facial marker information, and will instead prove that simple appearance-based techniques can achieve similar (comparable) results.

2.2 Facial Expression Recognition

The facial expression recognition component makes use of the NevenVision FaceTracker for facial feature point tracking, and produces 22 feature points as shown in the first sub-image of Figure 1. These points are used to locate the face, which is then cropped and normalized for accurate comparison. We adopt linear binary patterns (LBP) for the extraction and representation of facial expressions as they have been shown to be effective and efficient for emotion recognition, and also perform stably and robustly when fed compressed low-resolution video sequences [4]. The LBP operator assigns a code to every pixel of an image by thresholding the 3x3-neighbourhood of each pixel with the centre pixel value and considering the result as a binary number. This is represented by the equation below.

The local binary code for each pixel p in image I is

$$LBP_p = \sum_{n=0}^{N-1} s(g_n - g_p)2^n, \quad s(x) = \begin{cases} 1 & x >= 0 \\ 0 & x < 0 \end{cases} \qquad (1)$$

where n represents the neighbouring pixels ($N = 8$), g_n greyscale value of the neighbour pixel and g_p, greyscale value of centre pixel.

Once each pixel has been assigned a code, the image is divided into regions of 10x10 pixels to capture micro-patterns, such as edges and flat areas, that could help discriminate between the different facial expressions. Six-bin histograms of the LBP codes are computed for each region, and concatenated to form a feature vector which offers a concise representation of the face image (this method allows for spatial information to be retained).

For classification, the algorithm loops through a video sequence and obtains a feature vector for each frame. The feature vector is fed into Random Forests (with tree count of 100) from which average probabilities are calculated for every emotion. The probabilities are integrated over the time span of the video sequence to determine the most likely emotion.

2.3 Vocal Affect Analysis

Emotion can be detected from speech by analysing the characteristics of speech utterance waveforms. Voice cues such as the pitch, voice quality, intensity (perceived loudness) and temporal aspects of the speech indicate the nature of the tone or emotion behind an utterance. We employ an algorithm described in [6] which uses the OpenSMILE library [3] to extract low-level prosodic features such as the Mel-frequency cepstral coefficients, signal energy, and those mentioned above. Descriptive statistics (e.g. min, max, mean, percentiles, peaks, etc.) are computed and form part of the acoustic feature vector. The vector is then passed onto a correlation-based feature selection method to reduce dimensionality.

The classification task is achieved by training pairwise support-vector machines with radial basis function (RBF) kernels. The pairwise comparisons are combined through a voting mechanism; the output probabilities and the count of pairwise wins for each emotion are recorded for fusion purposes.

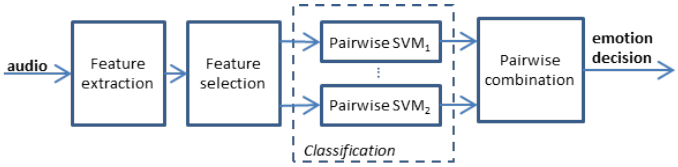

Figure 2: The pairwise framework for the inference of emotion from speech.

2.4 Multimodal Fusion

In Sharma's extensive introduction to fusion of multiple sensors [9], three distinct levels of integrating data are highlighted, namely, data, feature and decision fusion methods. Data fusion is automatically excluded from the consideration as it applies to observations of the same type (for example, two video camera recordings taken at different angles). Feature fusion is applied when the raw observations have been transformed into feature representations and is ideal for synchronized feeds. Decision fusion, also called late fusion, deals with the fusion of decisions computed independently by the respective components.

We have chosen decision fusion given its robust architecture and resistance to sensor failure. Although it has been noted that this approach loses information of mutual correlation between the audio and video modalities, it will nevertheless be sufficient for the database used in this study.

A linear SVM is used to combine the audio and video feeds. The probability estimates for each class obtained from the facial expression and vocal affect recognition components are combined to form a feature vector, which is fed into the SVM for classification.

3. EXPERIMENT & RESULTS

3.1 Experimental Setup

Two experiments with the following objectives were setup: (a) to compare the two modalities separately and integrated, (b) to study the effect of a corrupt signal on the overall performance of the system.

Figure 1: The extraction of linear binary patterns from an image to construct features for video emotion classification.

Table 1: Emotion classification accuracy (%) based on a five-fold cross validation

Emotion	Video classifier	Audio classifier	Bimodal classifier
Anger	100	83	98
Disgust	82	77	93
Fear	92	68	97
Happiness	98	70	98
Neutral	100	96	100
Sadness	95	72	100
Surprise	95	71	95
Weighted Average	95	79	98

The motive behind the second experiment is to synthesize a noisy environment as a step towards making HCI applications conform to our natural settings. The increasing trend of mobile computing highlights the necessity of HCI applications that will work with signals of poor quality. The experiment therefore aims to investigate the robustness of the fusion system when it is fed a corrupt or low resolution signal.

The first experiment was conducted using the five-fold cross-validation methodology. This resulted in each fold containing 96 files for testing and 380 files for training. The FaceTracker failed on two of the files which were then excluded from the experiments. One of the folds from the five partitions was used for analysis in the second experiment.

3.2 Results

3.2.1 Fusion Analysis

Table 1 lists average classification rates for the unimodal and bimodal classifiers over five folds. With the exception of the class *anger*, the audio-visual fusion model outperformed the individual modalities. These results are congruent with the findings presented in the *Related Works* section.

Table 2 shows a confusion matrix for the audio-visual fusion classifier. According to the results, *surprise* was misclassified with *happiness* three times. This could be explained by the common characteristics of the two emotions, that is, both having large mouth movements and being open expressions. The *anger, disgust* and *fear* emotions also have similar characteristics such as the wrinkling of the glabella and the nose region.

As mentioned earlier, the results above are discussed in relation with the work of Haq et al. The video-only classification rate reported was 98.3% and the fusion-classifier yielded a 98.3% recognition rate. The values obtained in our study, 95.2% and 98.0% for the respective categories, are comparable having relied solely on appearance-based methods.

Table 2: Confusion matrix of the bimodal classifier

Predicted / Actual [#]	A	D	F	H	N	Sa	Su
Anger	59	1	0	0	0	0	0
Disgust	1	56	2	0	1	0	0
Fear	0	0	57	0	0	2	0
Happiness	0	0	0	59	1	0	0
Neutral	0	0	0	0	120	0	0
Sadness	0	0	0	0	0	60	0
Surprise	0	0	0	3	0	0	56

The high video classification accuracy could be a reflection of the weak database chosen and not necessarily the strength of the LBP algorithm. The scarcity of fully-labelled, public audio-visual databases remains a problem in the field, especially those related to categorical emotions.The performance could also be attributed to the nature of the emotion set. Basic emotions are noted for carrying signature expressions (for example, nose wrinkle for *disgust* and brow furrows for *anger*) which make it easy for template matching.

3.2.2 Noise Analysis

A subset of videos were corrupted by adding white Gaussian noise. Figure 3 shows the result of the noise infusion at different noise variances. An increase in noise variance leads to difficulty in deciphering the images. These videos were passed through the emotion recognition system, with the sound untainted. The effect of the image noise on the system is depicted in Figure 4.

The figure shows the performance of the three classifiers as the Gaussian noise variance is increased from 0 to 0.005. The recognition capability of the facial expression subsystem decreases as the noise is increased, with the classification rate reaching 39.6% for a variance of 0.005. Although the fusion classification rate decreases with the variance, it does so at a significantly lower rate than that of the video classifier, and maintains an average that is greater than 80%.

This experiment proves that multimodal systems (and in particular decision-based fusion systems) are able to with-

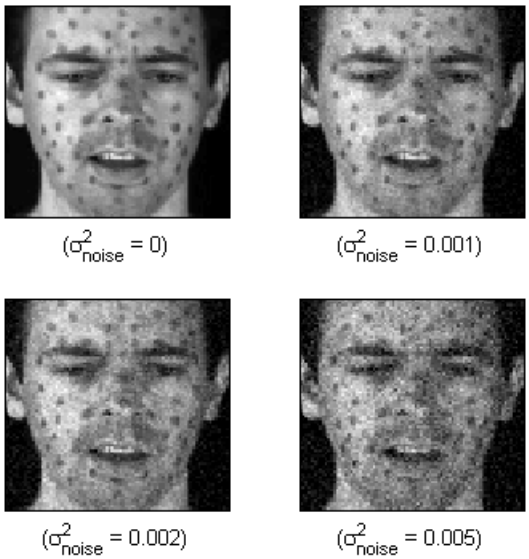

Figure 3: Samples of images corrupted with different variance levels of white Gaussian noise.

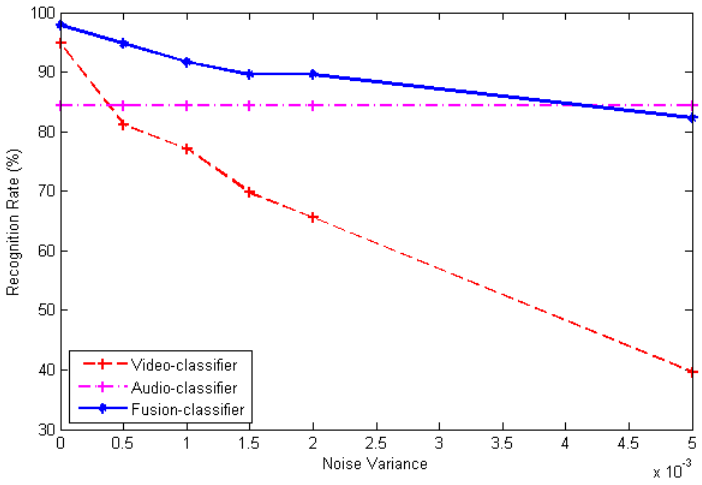

Figure 4: The peformance of the three classifiers under varying levels of white Gaussian noise.

stand sensor failures and are therefore ideal for environments which are susceptible to noise, and ideal for hardware with quality problems.

4. CONCLUSIONS

We investigated the fusion of video and audio channel for emotion recognition and the effect of image noise on the performance of the system. We found that the bimodal approach yields higher classification accuracies (97.7% compared to 95.25% and 79.1% for video-only and audio-only classifiers respectively) as a result of the complementary cues found in the different input channels.

Corrupting the video signal with noise showed that despite the poor performance of the video-only classifier (classification rate of 39.6%), the bimodal classifier was able to assign more weight to the audio classifier and therefore maintain a good overall classification rate. This confirmed the premise that decision-based multimodal systems are not compromised by sensor failures, and that they can be deployed in environments which are highly susceptible to noise.

Future work includes extending the multimodal fusion approach to a naturally-elicited complex emotion set consisting of mental states such as *thinking* and *concentrating*. This a good step towards designing for scenarios that are likely to happen in normal human-computer interactions. We plan to study the effect of noise on both the video and audio signals while moving toward realistic noisy conditions, such as varying lighting conditions and ambient sounds. This work will be applied to both early and late fusion techniques for a comparative analysis.

5. ACKNOWLEDGMENTS

The financial assistance of the South African National Research Fund (NRF), Bradlow Foundation Scholarship and Google Anita Borg Memorial Scholarship towards this research is hereby acknowledged.

6. REFERENCES

[1] S. Campanella and P. Belin. Integrating face and voice in person perception. *Trends in Cognitive Sciences*, 11(12):535–543, 2007.

[2] J. Ciarrochi, J. Forgas, and J. Mayer. *Emotional intelligence in everyday life: A scientific inquiry*. Psychology Pr, 2001.

[3] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the international conference on Multimedia*, pages 1459–1462. ACM, 2010.

[4] X. Feng, M. Pietikainen, and A. Hadid. Facial expression recognition based on local binary patterns. *Pattern Recognition and Image Analysis*, 17(4):592–598, 2007.

[5] S. Haq, P. Jackson, and J. Edge. Audio-visual feature selection and reduction for emotion classification. In *Proc. AVSP*, pages 185–190, 2008.

[6] T. Pfister and P. Robinson. Real-time recognition of affective states from non-verbal features of speech and its application for public speaking skill analysis. *IEEE Trans. Affective Computing (TAFFC)*, 2011.

[7] R. Picard. *Affective computing*. The MIT press, 2000.

[8] B. Schuller, D. Arsic, F. Wallhoff, and G. Rigoll. Emotion recognition in the noise applying large acoustic feature sets. *Speech Prosody, Dresden*, pages 276–289, 2006.

[9] R. Sharma, V. Pavlovic, and T. Huang. Toward multimodal human-computer interface. *Proceedings of the IEEE*, 86(5):853–869, 1998.

[10] M. You, C. Chen, J. Bu, J. Liu, and J. Tao. Emotion recognition from noisy speech. In *2006 IEEE International Conference on Multimedia and Expo*, pages 1653–1656. IEEE, 2006.