Comparison of Prediction-based Fusion and Feature-level Fusion Across Different Learning Models

Stavros Petridis
Department of Computing
Imperial College London
sp104@imperial.ac.uk

Sanjay Bilakhia
Department of Computing
Imperial College London
sb1006@imperial.ac.uk

Maja Pantic
Dept. of Computing / EEMCS
Imperial College London /
Univ. Twente
m.pantic@imperial.ac.uk

ABSTRACT

There is evidence in neuroscience indicating that prediction of spatial and temporal patterns in the brain plays a key role in perception. This has given rise to prediction-based fusion as a method of combining information from audio and visual modalities. Models are trained on a per-class basis, to learn the mapping from one feature-space to another. When presented with unseen data, each model predicts the respective feature-sets using its learnt mapping, and the prediction error is combined within each class. The model which best describes the audiovisual relationship (by having the lowest combined prediction error) provides its label to the input data. Previous studies have only used neural networks to evaluate this method of combining modalities - this paper extends this to other learning methods, including Long Short-Term Memory recurrent neural networks (LSTMs), Support Vector Machines (SVMs), Relevance Vector Machines (RVMs), and Gaussian Processes (GPs). Our results on cross-database experiments on nonlinguistic vocalisation recognition show that feature-prediction significantly outperforms feature-fusion for neural networks, LSTMs, and GPs, while performance on SVMs and RVMs is more ambiguous and neither model gains an absolute advantage over the other.

Categories and Subject Descriptors

I.5.4 [**Computing Methodologies**]: Pattern Recognition—*Applications*; J.m [**Computer Applications**]: Miscellaneous

General Terms

Algorithms, Experimentation

Keywords

Prediction-based Classification/Fusion, Audiovisual Fusion, Nonlinguistic Information Processing

MM'12, October 29–November 2, 2012, Nara, Japan.
Copyright 2012 ACM 978-1-4503-1089-5/12/10 ...$10.00.

1. INTRODUCTION

Audiovisual fusion approaches have been successfully applied to speech recognition [11], affect recognition [13] and very recently to laughter recognition [7, 9]. However, the optimal fusion type remains an open issue and largely depends on the problem.

The most common types of audiovisual fusion are feature-level fusion [12], where audio and visual feature vectors are concatenated and fed to some learning algorithm, and decision-level fusion [12], where each modality is modelled independently, and predicted labels are combined in some manner, using e.g. a linear sum rule, or a second level classifier. Recently, other types of fusion have also been proposed, like prediction-based fusion [7, 10].

Prediction-based fusion is based on the idea that the relationships between real-valued audio and visual features are different within each class. These relationships are therefore explicitly modelled, by learning the mapping from the audio-to-visual and visual-to-audio features, within each class, using some regression method. It is expected that the model which corresponds to the true class will produce a better prediction than all others. Class labels are predicted by selecting the model that produces the lowest error and most accurately describes the relationship between the two features sets for that class. The absolute error value for a prediction is unimportant - only its relative value compared to the predictions of other models is important.

In this study we compare the performance of prediction-based fusion with feature-level fusion on nonlinguistic vocalisations recognition. Our motivation is to investigate if the superior performance of prediction-based fusion over feature-level fusion using neural networks [7] generalises to other learning algorithms. We use five different algorithms that support both classification and regression: SVMs, RVMs, GPs, feedforward neural networks NNs and LSTMs. The performance is evaluated on cross-database experiments using 3 datasets, which poses a challenging test condition. Our results show that prediction-based fusion outperforms feature-level fusion in most cases (NNs, LSTMs and GPs), while for SVMs and RVMs neither classification model gains an absolute performance advantage over the other; either one method always performs worse than the other in at least one performance metric, or the difference in performance is not statistically significant.

2. DATABASES

For the purpose of this study we used three datasets which correspond to the different scenarios as explained below.

Table 1: Description of the datasets.

Type	No Episodes / No Subjects	Total Duration (sec)	Mean / Std (sec)
AMI			
Laughter	124 / 10	145.4	1.17 / 0.7
Speech	154 / 10	285.9	1.86 / 1.1
SAL - Training			
Laughter	57 / 10	80.6	1.4 / 0.8
Speech	96 / 10	204.3	2.1 / 0.8
SAL - Validation			
Laughter	37 / 5	50.3	1.4 / 0.7
Speech	81 / 5	159.3	2.0 / 0.8
MAHNOB			
Laughter	554 / 22	863.7	1.56 / 2.2
Speech	845 / 22	2430.9	2.88 / 2.3

AMI: In the AMI Meeting Corpus [5] people show a huge variety of spontaneous expressions. We only used the close-up video recordings of the subject's face and the corresponding audio recordings from individual subject headsets. The camera is fixed, and since people are involved in a multi-participant discussion, there is significant head movement. For our experiments we used seven meetings (IB4001 to IB4011) and the corresponding recordings of ten participants, 8 males and 2 females.

SAL: In the SAL database [2], the subjects interact with 4 different agents that have different personalities, and the audiovisual response of the subjects while interacting is recorded. For our experiments we used 15 subjects - 8 males and 7 females, out of which 10 are used for training and 5 for validation. We used the close-up video recordings of the subjects' face and the related audio recordings. Subjects have mostly frontal pose, and head movements are small.

MAHNOB: In the MAHNOB database [1, 8], laughter was elicited by showing amusing videos to subjects. There are 22 subjects in total - 12 males and 10 females. There is significant variation in the types of laughter recorded. The camera is fixed, and since subjects are watching a fixed screen, they are mostly in frontal pose. Two audio streams are available, from the camera microphone and from the lapel microphone. In this study, we used audio from the camera microphone only, since the data is noisier, and poses a more challenging generalisation problem.

All laughter and speech episodes used in this study were pre-segmented based on audio. This means that the start and end point of a laughter episode is pre-defined for the audio signal and then the corresponding video frames are extracted. For the AMI [5] and MAHNOB [1] datasets, laughter episodes were selected based on the annotations provided. For the SAL dataset, we manually annotated laughter episodes. Details of the four datasets are given in Table 1.

3. FEATURES

Audio Features: Cepstral features, such as MFCCs, have been widely used in speech recognition and have also been successfully used for laughter detection [4]. We use the first 6 MFCCs (together with 6 ΔMFCCs), given the findings in [4]. These are computed every 10ms, over a window of 40ms, making the frame rate 100 frames per second (fps).

Visual Features: Changes in facial expression are captured by tracking 20 facial points. These points are the corners of the eyebrows (2 points), the eyes (4 points), the nose (3 points), the mouth (4 points) and the chin (1 point) [6]. For each video segment containing K frames, we obtain a set of K vectors containing 2D coordinates of the 20 points. Using a Point Distribution Model (PDM), by applying principal component analysis to the matrix of these K vectors, head movement can be decoupled from facial expression. Using the approach proposed in [3], the facial expression movements are encoded by the projection of the tracking points coordinates to the N principal components (PCs) of the PDM which correspond to facial expressions. For the SAL dataset, it was found that 3 PCs encode mostly facial expressions (PCs 5,6,7). Further details of the feature extraction procedure can be found in [9, 3].

4. PREDICTION-BASED FUSION

For each class (speech, laughter) we train two regression models - one model learns the mapping from audio to visual features, one model learns the mapping from visual to audio features.

So, the first model takes an audio feature-vector as input, and gives its prediction for the corresponding visual feature-vector at the same frame, as output. The second model takes a visual feature-vector as input, and gives its prediction for the corresponding audio feature-vector at the same frame, as output. Therefore the relationship between the audio (A^L, A^S) and visual (V^L, V^S) features in speech and laughter is modelled by (f^L_{AV}), (f^L_{VA}) for laughter and (f^S_{AV}), (f^S_{VA}) for speech.

$$f^L_{AV}(A^L) = \hat{V}^L \approx V^L \tag{1}$$
$$f^L_{VA}(V^L) = \hat{A}^L \approx A^L \tag{2}$$
$$f^S_{AV}(A^S) = \hat{V}^S \approx V^S \tag{3}$$
$$f^S_{VA}(V^S) = \hat{A}^S \approx A^S \tag{4}$$

Once the mappings (f^L, f^S) are learnt, an unseen example is classified using the pair of models that produce the lowest prediction error. When a new frame is available the audio and visual features are computed, and are then fed to the models from eq. 1 - 4, and 4 error values are produced, eq. 5 - 8. We use mean squared error (MSE). We then compute the combined MSE for each class, which takes into account the bidirectional relationship of audio and visual features, as shown in eq. 9 and 10, where w is a weighting factor.

$$e^L_{AV} = MSE(\hat{V}^L, V^L) \tag{5}$$
$$e^L_{VA} = MSE(\hat{A}^L, A^L) \tag{6}$$
$$e^S_{AV} = MSE(\hat{V}^S, V^S) \tag{7}$$
$$e^S_{VA} = MSE(\hat{A}^S, A^S) \tag{8}$$
$$e^L = w_L \times e^L_{AV} + (1 - w_L) \times e^L_{VA} \tag{9}$$
$$e^S = w_S \times e^S_{AV} + (1 - w_S) \times e^S_{VA} \tag{10}$$

A frame is labelled as laughter or speech depending on which pair of models (corresponding to a particular class) produced the best feature reconstruction, i.e. the pair with the lowest combined prediction error, eq. 9 and 10. In other words, a frame is labelled based on the following rule:

$$IF\ \ e^S > e^L\ \ THEN\ \ \mathbf{L}\ \ ELSE\ \ \mathbf{S} \tag{11}$$

5. EXPERIMENTAL STUDIES

In order to assess the performance of the method presented in section 4, cross database experiments between SAL, AMI and MAHNOB were performed. We performed experiments using feedforward NNs, LSTMs, SVMs, RVMs, and GPs, using both feature-level fusion and prediction-based fusion to combine modalities. The Matlab neural network toolbox, Pybrain, LIBSVM, SparseBayes, and GPML packages were used for each of learning algorithms, respectively.

5.1 Experimental Setup

Preprocessing: As mentioned in section 3, we used 3 visual features and 12 audio features in our experiments. Before training, the audio and visual features are synchronised by upsampling the visual features (using linear interpolation), to match the frame rate of the audio features (100fps). All the audio and visual features are z-normalized per subject in order to remove subject and recording variability.

Training: For training feedforward NNs, LSTMs and SVMs the first ten subjects of the SAL database are used, amounting to 28488 samples. GPs and RVMs compute a distance matrix between every pair of samples during training; this matrix is subsequently inverted. This operation is impractical given 28488 samples, and the distance matrix is likely near-singular, precluding inversion. We therefore randomly subsample, selecting 2000 samples from the pool of the first 10 subjects of SAL, and run the experiment 10 times, reporting mean and standard deviation.

Laughter and speech feature-prediction models are only trained with samples from their respective classes, while binary classifiers using feature fusion are trained on samples from both classes.

Parameter Optimization: Each underlying learning algorithm has its own unique set of parameters, whose values significantly affect performance. These are optimized as follows:

NNs, LSTMs: In both feedforward and LSTM cases, we trained networks with only one hidden layer. The size of this hidden layer (number of hidden neurons) was optimized using a line search across the range [5-30], in steps of 5 for feedforward networks, and the range of [10-90] in steps of 10 for LSTM networks. LSTM and feedforward networks were trained using resilient backpropagation, with a training epoch limit of 500.

SVMs: For binary classification, SVMs require two parameters to be optimized - the kernel width and the soft-margin cost, while support vector regression also takes ϵ, the width of the loss-insensitive region in the loss function. In both cases, these are found using a single-resolution gridsearch. Kernel width was optimized across the range 0.05 to 0.8, soft-margin cost 0.1 to 8, and ϵ 0.1 to 1.1.

RVMs: RVMs require only the kernel width to be optimized, for which we used a line search using the range 0.05 to 1.5.

GPs: Gaussian processes require only the kernel width to be optimized, for which we used a line search using the range 0.05 to 1.5.

In addition, feature prediction also requires optimization of the weights w_L and w_S from eq. 9, 10, with regard to classification performance. This is performed using a single-resolution gridsearch in steps of 0.05, subject to $0 < \text{w} < 1$.

For each point in the parameter space (for a particular learning algorithm), we train a classifier, using either feature-fusion or feature-prediction as a classification model. This is then evaluated on a validation set, consisting of the remaining 5 subjects of SAL. Both the feature-fusion and feature-prediction methods produce a class label per frame - we use majority voting over an entire sequence to convert this "bag" of frame labels to a sequence label. The best performing model is chosen (using f1 performance as a selection criterion), and tested as below.

Table 2: F1 and Unweighted Average Recall Rates (UAR) for the feature-level fusion (FF) system and the prediction based system tested on the AMI dataset.

Classification System	F1 Laughter	F1 Speech	UAR
NN			
A + V (F F)	66.7 (2.5)	81.3 (1.0)	73.9 (1.6)
A + V Pred.	73.3 (1.1)	84.2 (0.5)	78.3 (0.8)
LSTM			
A + V (F F)	54.3 (6.9)	79.2 (1.5)	68.3 (3.4)
A + V Pred.	68.6 (3.7)	82.5 (1.3)	75.4 (2.9)
SVM			
A + V (F F)	71.6 (0.0)	83.9 (0.0)	77.4 (0.0)
A + V Pred.	72.4 (0.0)	78.0 (0.0)	75.2 (0.0)
RVM			
A + V (F F)	71.9 (4.5)	82.8 (1.7)	77.1 (6.3)
A + V Pred.	72.3 (3.6)	82.1 (1.8)	76.9 (4.5)
GP			
A + V (F F)	49.5 (12.5)	16.5 (12.9)	40.1 (16.1)
A + V Pred.	77.8 (2.1)	85.5 (0.9)	81.3 (2.6)

Testing: SVMs, GPs, and RVMs are all deterministic methods, hence are retrained and tested once. LSTM and feedforward NNs are trained using resilient backpropagation, which gives random initial values to each of the network weights (hence each training process produces a different model); therefore we train and test 10 times, reporting mean and standard deviation. The performance measures we use are F1-score and Unweighted Average Recall (UAR) rates.

5.2 Results

Table 2 shows the experimental results on the AMI dataset. We see that feature prediction outperforms feature-fusion on both feedforward neural networks and LSTM networks, across all three measures of performance - laughter F1, speech F1, and UAR. Unusually, even though LSTMs take temporal information into account, we see that they perform worse than static feedforward networks - this may be due to Pybrain's inability to specify as many stopping criteria as the Matlab NNs toolbox (such as minimum gradient), possibly leading to overfitting. Feature-prediction using SVR is outperformed by feature-fusion on both speech F1 and UAR, by 5.9% and 2.2% respectively, though it does beat feature-fusion on laughter F1, suggesting that SVMs tend to favour speech, which is the class with the most examples. There appears to be no statistically significant difference in performance between feature-prediction and feature-fusion using relevance-vector machines for any performance measure. Gaussian process-based feature-prediction dramatically out-

Table 3: F1 and Unweighted Average Recall Rates (UAR) for the feature-level fusion (FF) system and the prediction based system tested on the MAHNOB dataset.

Classification	F1	F1	UAR
System	Laughter	Speech	
NN			
A + V (F F)	60.5 (1.1)	81.4 (0.4)	70.3 (0.6)
A + V Pred.	73.0 (1.7)	85.2 (0.7)	78.2 (1.2)
LSTM			
A + V (F F)	54.6 (7.8)	81.8 (1.8)	68.2 (4.5)
A + V Pred.	69.9 (2.4)	84.5 (0.7)	81.8 (2.7)
SVM			
A + V (F F)	60.4 (0.0)	82.5 (0.0)	70.7 (0.0)
A + V Pred.	65.0 (0.0)	73.8 (0.0)	70.1 (0.0)
RVM			
A + V (F F)	70.4 (4.3)	83.7 (1.5)	83.6 (6.8)
A + V Pred.	68.3 (3.4)	81.0 (1.7)	74.4 (6.3)
GP			
A + V (F F)	41.9 (15.8)	18.5 (25.1)	36.8 (18.7)
A + V Pred.	80.1 (1.3)	88.2 (0.6)	83.5 (2.2)

performs feature-fusion, by 28.3% on laughter-F1, 69% on speech-F1, and 41.2% on UAR, with relatively small standard deviation.

In Table 3 we can see the experiment results on the MAHNOB database. As with AMI, there is a statistically significant difference between the performance of feature-prediction over feature-fusion when using feedforward networks and LSTM networks, although feedforward networks once again outperform LSTM networks on both feature-prediction and feature-fusion. As when tested on AMI, SVR-based feature-prediction is outperformed by feature-fusion on speech-F1 and UAR, with the converse applying for laughter-F1. Prediction based on GPs once again significantly outperforms feature-fusion, despite the fact that the standard deviation for feature-fusion is very high; once again, GP-based prediction has an extremely small standard deviation. This suggests that Gaussian processes are sensitive to the size and quality of training data. The relative performance of RVMs on the MAHNOB dataset, however, is quite different from the performances on the AMI dataset. We see that feature-fusion for binary classification outperforms feature-prediction by a statistically significant amount for both speech-F1 and UAR; for laughter-F1 feature-fusion also outperforms feature-prediction, however the result is not statistically significant due to the high standard deviation of both methods.

6. CONCLUSIONS

We have compared the prediction-based fusion approach to feature-level fusion across different statistical models, in order to analyse the performance characteristics. Models based on NNs, LSTMs, SVMs, RVMs, and GPs were evaluated in cross-database experiments, to determine generalization capability when faced with testing data thats differs significantly from the training data. The key idea is that class predictions are made based on the model that best describes the spatial relationship between the audio and visual features, rather than their absolute values in the feature space. This should provide more robust performance. We find that for feedforward NNs, LSTMs and GPs, feature-prediction outperforms feature-level fusion, while for SVMs and RVMs, prediction-based fusion only outperforms feature-fusion in some cases, and performance depends on the quality of the training data.

7. ACKNOWLEDGMENTS

The research presented in this paper has been funded by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB).

8. REFERENCES

[1] http://mahnob-db.eu/laughter/.

[2] E. Douglas-Cowie, R. Cowie, C. Cox, N. Amir, and D. Heylen. The Sensitive Artificial Listener: an induction technique for generating emotionally coloured conversation. In *Workshop on Corpora for Research on Emotion and Affect*, pages 1–4.

[3] D. Gonzalez-Jimenez and J. L. Alba-Castro. Toward pose-invariant 2-d face recognition through point distribution models and facial symmetry. *IEEE Trans. Inform. Forensics and Security*, 2(3):413–429, 2007.

[4] L. Kennedy and D. Ellis. Laughter detection in meetings. In *NIST Meeting Recognition Workshop*, 2004.

[5] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, and V. Karaiskos. The AMI meeting corpus. In *Int'l. Conf. on Methods and Techniques in Behavioral Research*, pages 137–140, 2005.

[6] I. Patras and M. Pantic. Particle filtering with factorized likelihoods for tracking facial features. In *FG*, pages 97–104, 2004.

[7] S. Petridis, A. Asghar, and M. Pantic. Classifying laughter and speech using audio-visual feature prediction. In *IEEE ICASSP*, pages 5254–5257, 2010.

[8] S. Petridis, B. Martinez, and M. Pantic. The MAHNOB laughter database. In *Submitted to Image and Vision Computing Journal*, 2012.

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

[10] S. Petridis, M. Pantic, and J. Cohn. Prediction-based classification for audiovisual discrimination between laughter and speech. In *IEEE FG*, pages 619–626, Santa Barbara, CA, USA, 2011.

[11] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior. Recent advances in the automatic recognition of audiovisual speech. *Proc. of the IEEE*, 91(9):1306–1326, 2003.

[12] C. Snoek, M. Worring, and A. Smeulders. Early versus late fusion in semantic video analysis. In *ACM Multimedia*, pages 399–402, 2005.

[13] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A survey of affect recognition methods: Audio, visual and spontaneous expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.