Bimodal Log-linear Regression for Fusion of Audio and Visual Features

Ognjen Rudovic
Dept. of Computing
Imperial College London
London, UK
o.rudovic@imperial.ac.uk

Stavros Petridis
Dept. Computing
Imperial College London
London, UK
sp104@imperial.ac.uk

Maja Pantic
Dept. Computing / EEMCS
Imperial College London /
Univ. Twente
London, UK / Enschede, NL
m.pantic@imperial.ac.uk

ABSTRACT

One of the most commonly used audiovisual fusion approaches is feature-level fusion where the audio and visual features are concatenated. Although this approach has been successfully used in several applications, it does not take into account interactions between the features, which can be a problem when one and/or both modalities have noisy features. In this paper, we investigate whether feature fusion based on explicit modelling of interactions between audio and visual features can enhance the performance of the classifier that performs feature fusion using simple concatenation of the audio-visual features. To this end, we propose a log-linear model, named Bimodal Log-linear regression, which accounts for interactions between the features of the two modalities. The performance of the target classifiers is measured in the task of laughter-vs-speech discrimination, since both laughter and speech are naturally audiovisual events. Our experiments on the MAHNOB laughter database suggest that feature fusion based on explicit modelling of interactions between the audio-visual features leads to an improvement of 3% over the standard feature concatenation approach, when log-linear model is used as the base classifier. Finally, the most and least influential features can be easily identified by observing their interactions.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation

Keywords

Bimodalss Log-Linear Regression-based Fusion, Audiovisual Fusion, Laughter Classification

MM'13, October 21–25, 2013, Barcelona, Spain.
Copyright 2013 ACM 978-1-4503-2404-5/13/10 ...$15.00.
http://dx.doi.org/10.1145/2502081.2502207 .

1. INTRODUCTION

Audiovisual fusion has attracted significant interest given its successful application in speech recognition, affect recognition and lately in laughter recognition [9]. The main contribution of the visual information is the addition of complementary and redundant information which cannot be corrupted by acoustic noise and therefore may improve the performance of a recognition system.

The most common types of audiovisual fusion are decision-level and feature-level fusion [10]. In decision-level fusion the audio and video modalities are processed independently and then the classifiers outputs are combined using various integration rules, e.g. a linear sum rule, or a second level classifier. As a consequence, the correlation between the audio and visual features is lost. In feature-level fusion the extracted audio and visual features are first combined, usually through concatenation, and then fed to a classifier. This increases the dimensionality of the problem but it provides additional information that can be relevant for a target task. In this paper, we limit our consideration to feature-level fusion.

The goal of this study is to investigate whether explicit modelling of interactions between audio and visual features can enhance performance of the traditional feature fusion approach where audio and visual features are simply concatenated. To this end, we propose a log-linear model, named Bimodal Log-linear regression, that modifies the parametrisation of the logistic regression model for binary outputs in order to account for interactions between audio and visual features. We should point out that other approaches which take into account the interactions between audio and visual features exist in the literature like [2, 5, 7]. However, the aim of this work is to compare the performance of the Bimodal Log-Linear Regression with the standard logistic regression.

The effectiveness of the proposed classifier is evaluated in the task of laughter-vs-speech discrimination, since both laughter and speech are naturally audiovisual events. Finally, the experimental evaluation is conducted using the recently released MAHNOB laughter database [8], which contains hundreds of examples of audiovisual speech and laughter episodes.

2. DATASET AND EXTRACTED FEATURES

For the purpose of this study we used the MAHNOB Laughter audiovisual dataset [1, 8]. In this dataset, laughter was elicited by showing amusing videos to subjects. In a

Table 1: Description of the MAHNOB dataset.

MAHNOB			
Type	No Episodes /	Total Duration	Mean / Std
	No Subjects	(sec)	(sec)
Laughter	554 / 22	863.7	1.56 / 2.2
Speech	845 / 22	2430.9	2.88 / 2.3

different set of sessions subjects were asked either to discuss with a friend or talk about a topic of their choice in English. In all cases, their reactions were recorded by a fixed camera (720 x 576, 25 fps), and two microphones, the camera microphone and a lapel microphone. Given that the subjects are watching a fixed screen, they are mostly in frontal pose and there is not significant head movement. In total, there are 22 subjects in total, 12 males and 10 females, with all but one being non-native speakers. In this study, we used the audio from the camera microphone (48 kHz) only, since it is noisier, and poses a more challenging generalisation problem. Finally, we use the annotations for speech and laughter provided with the database. The details of the dataset are summarised in Table 1.

Audio Features: In this study we use the most commonly used features in speech processing, the Mel Frequency Cepstral Coefficients (MFCCs). We use the first 6 coefficients, together with their deltas, which capture some local temporal characteristics. So in total there are 12 features which are computed every 10ms over a window of 40ms, i.e. the frame rate is 100 fps.

Visual Features: To capture face movements in an input video, we track 20 facial points using the particle filtering algorithm proposed in [6]. These points are the corners/extremities of the eyebrows (4 points), the eyes (8 points), the nose (3 points), the mouth (4 points) and the chin (1 point). The features are computed using the same Point Distribution Model (PDM) as in [9]. As suggested in [4], 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. PCs 7-10 were found to correspond to facial expressions so our shape features are the projection of the 20 points to those 4 PCs. Further details of the feature extraction procedure can be found in [9, 4].

3. METHOD

In this section we present two different feature fusion methods for classification of speech and laughter. Throughout the paper, we assume a supervised setting: we are given a training set of N data triplets $D = \{(\mathbf{x}_i^A, \mathbf{x}_i^V, \mathbf{y}_i)\}_{i=1}^N$, which are i.i.d. samples from an underlying but unknown distribution $p_*(\mathbf{y}, \mathbf{x})$. Furthermore, $\mathbf{x}_i^A = (x_1^A ... x_{T_i}^A)^T$ and $\mathbf{x}_i^V = (x_1^V ... x_{T_i}^V)^T$ contain audio and visual features, respectively, per sample, and $\mathbf{y}_i = \{0, 1\}$ is the class label per sequence, with 0 denoting the speech class, and 1 the laughter class.

3.1 Logistic Regression

Logistic regression [3] is the simplest form of the log-linear models that deal with binary outputs. This model is appealing for the target task (i.e., classification of speech and laughter) because it is a discriminative model, so it directly models what we want, $p(y|x)$. Its predictor variables x can take any form since, in contrast to generative models, this model makes no assumption about the distribution of x. Therefore, x does not have to be normally distributed, linearly related or of equal variance within each group. Formally, logistic regression models the class-conditional probability given by

$$p(y = 1|x, \beta) = \frac{1}{1 + \exp(-s(x, \beta))} \tag{1}$$

and $p(y = 0|x, \beta) = 1 - p(y = 1|x, \beta)$, where the feature function $s(x, \beta)$ is defined as $s(x, \beta) = \beta_0 + \sum_{j=1}^{d} \beta_j x_j$. The parameter vector β has as its elements the intercept β_0 and the weights β_j, where we use j to index over the feature values x_1 to x_d of a single example of dimensionality d.

In the context of the target task, logistic regression can be used to perform feature fusion by concatenating audio and visual features per sample, i.e., $x = (x^A, x^V)$. Correspondingly, the score function of the logistic regression can be written as

$$\begin{aligned} s(x^A, x^V, \beta) &= \beta_0 + \textstyle\sum_{j=1}^{d_A} \beta_j^A x_j^A + \sum_{j=1}^{d_V} \beta_j^V x_j^V \\ &= \beta \cdot \left[1 \ x^A \ x^V\right]^T \end{aligned} \quad , \tag{2}$$

where $\beta = \{\beta_0, \beta^A, \beta^V\}$, and d^A and d^V are the dimensions of the audio and visual features, respectively. With such feature function, the logistic regression is used to obtain the class-predictions per sample, which are then combined to obtain the class-probability for the whole sequence. This is explained in the following section.

3.2 Bimodal Log-linear Regression

Logistic regression [3] is limited in that its feature function treats audio and visual features independently. Thus, it ignores interactions between different audio and visual features that could be important for the classification task. To account for feature-interaction between the two modalities, we employ the log-linear model, which is an extension of logistic regression that allows modelling of arbitrary relationships among the input features. Specifically, we define the score function of the log-linear model as

$$s(x^A, x^V, \beta) = \beta_0 + \sum_{j=1}^{d_A} \beta_j^A x_j^A + \sum_{j=1}^{d_V} \beta_j^V x_j^V + \sum_{j=1}^{d_A} \sum_{k=1}^{d_V} \beta_{jk}^{AV} x_j^A x_k^V , \tag{3}$$

where β^{AV} measures the relevance of interactions between different audio and visual features for the classification task. Note that since we are interested in the fusion task we model only the first-order interactions between the features of two modalities, but not the features within these modalities. However, these and higher order interactions can be easily incorporated in the score function of the model. By plugging the score function in (3) into the conditional model in (1), we obtain a log-linear model for combining two modalities, and which we name Bimodal Log-linear Regression (BLR).

To better understand the BLR model, we re-write the score function in (3) using the matrix form

$$s(x^A, x^V, \beta) = \begin{bmatrix} 1 & x^A \end{bmatrix} \cdot \beta \cdot \begin{bmatrix} 1 \\ (x^V)^T \end{bmatrix} . \tag{4}$$

By comparing the score functions in (2) and (3), we see that the rank of the parameter vector β for logistic regression is 1, whereas for BLR is $r = min\{d_A, d_V\}$. Therefore, the former model performs linear classification in 1-D space, and

Table 2: The performance of the models evaluated on the MAHNOB dataset. L-F1 and S-F1 are obtained F-1 measures for Laughter and Speech, respectively, and CR is the classification rate, computed using all test subjects.

Model	L-F1[%]	S-F1[%]	CR[%]
A-LR	79.3	87.1	84.7
V-LR	82.5	88.6	85.9
AV-LR	87.2	91.1	89.4
AV-BLR	**90.5**	**94.3**	**92.7**
AV-NN[8]	86.5	92.2	90.1

the latter does so in the r-D space, in which the target classes may be better separated.

The training of the BLR model is accomplished by maximising the (balanced) conditional log-likelihood

$$\mathcal{L}(\beta) = \sum_{k=0}^{1} \frac{1}{N_k} \sum_{i=1}^{N_k} \sum_{j=1}^{T_i} \log p(y_i = k | x_{ij}^A, x_{ij}^V, \beta) - \lambda \, \|\beta\|_2^2 \, , \tag{5}$$

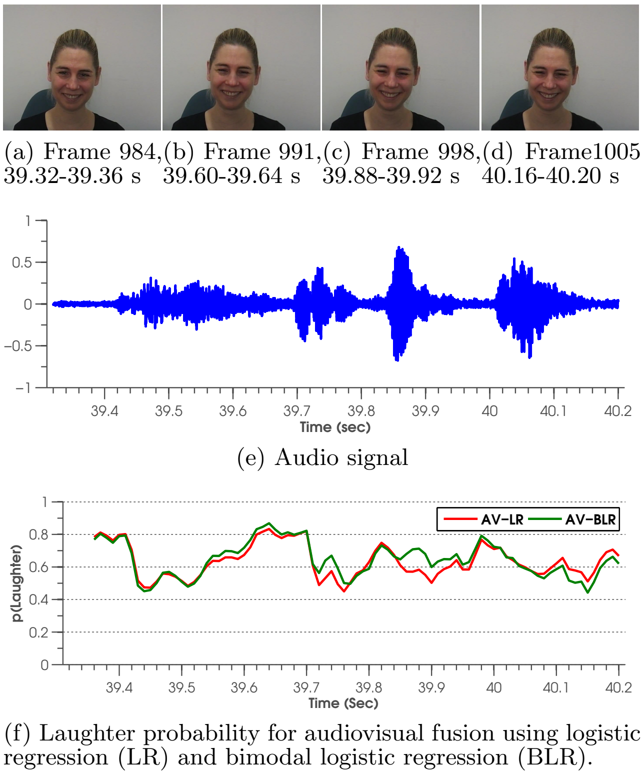
where N_k is the number of sequences of the k-th class, and T_i is the number of frames in the i-th sequence. Parameter regularisation is attained using the Frobenious norm, which implicitly enforces low-rank representation of β. We also balance the data likelihood based on the number of samples from each class. This is important since the speech class usually dominates the laughter class in the number of training examples. The conditional log-likelihood in (5) is strictly convex, so we use gradient ascent with a fixed step γ to find the (globally) optimal parameters β as

$$\beta_{t+1} = \beta_t + \gamma \nabla \mathcal{L}(\beta_t), \tag{6}$$

where t is the number of iterations and $\gamma = 0.1$. In all our experiments, the algorithm converged in less than 10 iterations. Finally, inference of test sequences is performed using a two-step approach: first, we perform classification of the input features per frame based on the probability in (1), where the winning class is the class with $p > 0.5$. Then, we apply majority voting to obtain the sequence label.

4. EXPERIMENTS

We test the performance of the proposed BLR model in the task of laughter-vs-speech discrimination. In all our experiments, we applied a leave-one-subject-out cross-validation procedure. The regularisation term is obtained by running another cross validation among $\lambda = \{10^{-5}, 10^{-4}, ..., 1, 2\}$ on the training set. Both audio and visual features are z-normalized to zero mean and unity standard deviation. The means and standard deviations are computed on the training set only, and then applied to the test set. In addition, the audio and visual features are synchronised for audiovisual fusion, since they are extracted at different frame rates. This is achieved by upsampling the visual features by linear interpolation as in [9]. The generalisation performance is reported using F-1 measure (per class) and the classification rate (CR), computed on the predicted labels for sequences of all test subjects. We compare the BLR model, trained using audio-visual (AV) features, to logistic regression (LR) models trained using: (i) audio (A-LR), (ii) visual (V-LR), and (iii) concatenated audio-visual (AV-LR) features, as explained in Sec. 3.1. Table 2 shows comparative results of the tested models. The models that rely on the single modality (A-LR and V-LR) are outperformed by the feature-fusion-based models. This is expected and the reader is referred to [8] for more insights about this. The proposed AV-BLR model attains better performance in terms of both F-1 measure and CR than the AV-LR model. We account this to modelling of the interactions between the two modalities in the BLR model. For the sake of comparison, we also include the baseline results for the MAHNOB dataset [8] obtained using a feedforward neural network for audio-visual feature fusion (AV-NN).



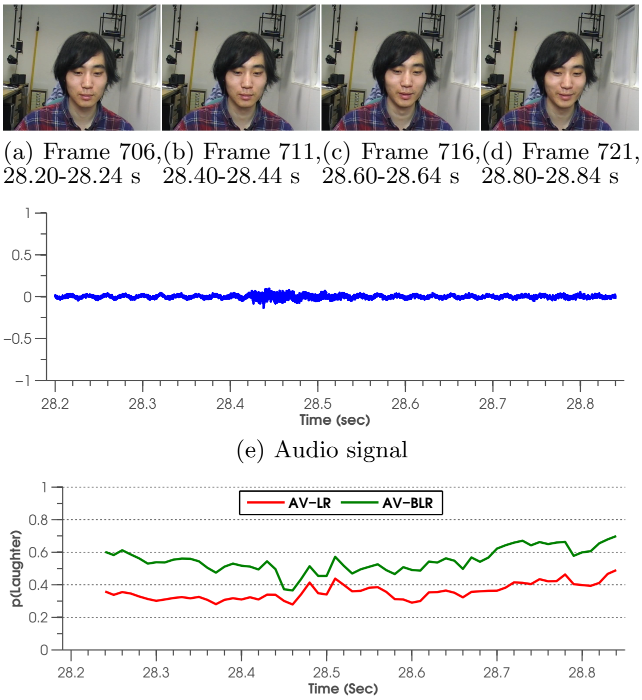
(a) Frame 984, 39.32-39.36 s (b) Frame 991, 39.60-39.64 s (c) Frame 998, 39.88-39.92 s (d) Frame1005 40.16-40.20 s

(e) Audio signal

(f) Laughter probability for audiovisual fusion using logistic regression (LR) and bimodal logistic regression (BLR).

Figure 1: Example from subject S009, S009-001.

To better understand the effect of modelling the interactions between the audio and visual features in the AV-BLR model, in Fig.1 and 2 we contrast its classification per frame to that attained by the AV-LR model on two example sequences of Laughter. Note from Fig.1 that in the case of ‘pronounced’ laughter, both models achieve similar performance, with the AV-BLR model slightly outperforming the AV-LR model. This is expected since both audio and visual features vary sufficiently to discriminate between the target classes. However, note that in the case of ‘soft’ laughter (Fig.2), it is difficult to extract from the noisy audio signal the features that are typical for laughter. Although the probabilities of both models lie in close vicinity of the classification margin (p=0.5), the proposed AV-BLR model exploits the correlations between the two modalities, which, evidently, helps to extract useful information from the noisy audio features that are highly correlated with the visual features. This is especially the case in the second half of the sequence in Fig.2, where visual features provide more evidence of laughter, and hence, amplify the influence of the correlated audio features.



(a) Frame 706, 28.20-28.24 s (b) Frame 711, 28.40-28.44 s (c) Frame 716, 28.60-28.64 s (d) Frame 721, 28.80-28.84 s

(e) Audio signal

(f) Laughter probability for audiovisual fusion using logistic regression (LR) and bimodal logistic regression (BLR).

Figure 2: Example from subject S005, S005-008.

Finally, we inspect the parameters β of the AV-BLR model in order to see to what extent AV-BLR was able to learn interactions between the two modalities. To this end, we visualise β in Fig. 3. Note that the first visual feature has the highest influence on the score function in the AV-BLR model. The first visual feature corresponds to the 7th PC that captures mouth movements that are important for discriminating laughter vs. speech. Also, the first six MFCC features have much more influence than their deltas (features 8 to 13). Finally, note that the model has learned different weights for correlations between audio and visual features. For example, from Fig. 3, we can see that the first visual feature is negatively correlated with 1-3 MFCC features. These correlations make the whole classification process more robust to noise in the features of both modalities, as evidenced by the experiments presented before.

5. CONCLUSION

We have proposed a new log-linear model for bimodal feature fusion, named Bimodal Log-linear regression. In contrast to the standard feature-level fusion that is based on a simple feature concatenation, BLR explicitly models interactions between two modalities. Our experiments conducted on the MAHNOB dataset show that modelling those interactions improves the discriminative power of the log-linear classifier. In addition, the most influential features and their interactions can be easily identified by visualizing the fusion matrix of the BLR model.

6. 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). The work of Stavros Petridis is also supported in part by EPSRC grant EP/H016988/1: Pain rehabilitation: E/Motion-based automated coaching.

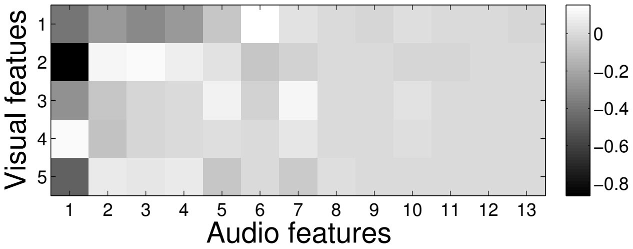


Figure 3: β parameters of AV-BLR. The element (1,1) corresponds to the intercept (β_0) of the model in (3). The rest of the elements in the first row and column are the weights of audio (β^A) and visual (β^V) features. The remaining elements are the weights of the interactions between the audio and visual features (β^{AV}). The (absolute) value of each element shows how strong the influence of the feature (feature pairs) on the score function in AV-BLR, and, thus, on its classification decision is.

7. REFERENCES

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

[2] Z. Barzelay and Y. Schechner. Harmony in motion. In *IEEE CVPR*, pages 1–8, 2007.

[3] D. Cho and T. Bui. Multivariate statistical modeling for image denoising using wavelet transforms. *Signal Processing: Image Comm.*, 20(1):77–89, Jan. 2005.

[4] 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.

[5] W. Jiang and A. C. Loui. Audio-visual grouplet: temporal audio-visual interactions for general video concept classification. In *ACM MM*, pages 123–132, New York, NY, USA, 2011.

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

[7] S. Petridis, S. Bilakhia, and M. Pantic. Comparison of prediction-based fusion and feature-level fusion across different learning models. In *ACM Multimedia 2012*, pages 813–816, Nara, Japan, November 2012.

[8] S. Petridis, B. Martinez, and M. Pantic. The MAHNOB laughter database. *Image and Vision Computing Journal*, 31(2):186–202, February 2013.

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