Audio-visual Classification and Fusion of Spontaneous Affective Data in Likelihood Space

Mihalis A. Nicolaou*, Hatice Gunes* and Maja Pantic*,†
*Department of Computing, Imperial College London, U.K.
†Faculty of EEMCS, University of Twente, The Netherlands
(michael.nicolaou08, h.gunes,m.pantic)@imperial.ac.uk

Abstract

This paper focuses on audio-visual (using facial expression, shoulder and audio cues) classification of spontaneous affect, utilising generative models for classification (i) in terms of Maximum Likelihood Classification with the assumption that the generative model structure in the classifier is correct, and (ii) Likelihood Space Classification with the assumption that the generative model structure in the classifier may be incorrect, and therefore, the classification performance can be improved by projecting the results of generative classifiers onto likelihood space, and then using discriminative classifiers. Experiments are conducted by utilising Hidden Markov Models for single cue classification, and 2 and 3-chain coupled Hidden Markov Models for fusing multiple cues and modalities. For discriminative classification, we utilise Support Vector Machines. Results show that Likelihood Space Classification improves the performance (91.76%) of Maximum Likelihood Classification (79.1%). Thereafter, we introduce the concept of fusion in the likelihood space, which is shown to outperform the typically used model-level fusion, attaining a classification accuracy of 94.01% and further improving all previous results.

1. Introduction

Human communicative modalities are multiple and not occurring in predetermined, restricted and controlled settings. Mainstream research on automatic affect sensing and recognition has focused on recognition of facial and vocal expressions in terms of basic emotional states (neutral, happiness, sadness, surprise, fear, anger and disgust), and based on data that has been posed on demand or acquired in laboratory settings [5, 17]. However, a number of researchers have shown that in everyday interactions people exhibit non-basic and subtle affective states, expressed via dozens (possibly hundreds) of anatomically possible facial expressions and bodily gestures, or linguistic and paralinguistic messages. These researchers advocate the use of dimensional description of human affect, where an affective state is characterised in terms of a number of latent dimensions [15]. According to this approach, the majority of affective variability is covered by two dimensions: valence (V, how positive or negative the emotion is) and arousal (A, how excited or apathetic the emotion is) [8].

When applying the aforementioned approach to automatic dimensional affect recognition, a common methodology is to reduce the classification problem to a two-class problem (positive vs. negative and active vs. passive classification problem) or to a four-class problem (classification into the quadrants of 2D A-V space). [1] use feedforward back-propagation networks for mapping into neutral and A-V quadrants. [16] work with the audio channel of SAL database and quantise the A-V into 4 or 7 levels and use Conditional Random Fields to predict the quantised labellings. Details on the aforementioned works, and an overview of the current efforts in the field of automatic dimensional affect recognition can be found in [5].

The work introduced in this paper is aligned with the recent shift in the field by being the first approach to focus on automatic recognition of spontaneous affect from facial, shoulder and audio cues in terms of *discretised* descriptions in the valence dimension. Our goal is to analyse audio-visual segments portraying spontaneous emotional expressions either as negative or positive by utilising coupled Hidden Markov Models ((C)HMMs). As (C)HMMs are generative models, separate models are trained for each class. Given an observation sequence, each model then outputs the likelihood of itself having generated the observation at hand.

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.900

The standard rule, commonly applied in previous works in automatic affect recognition (e.g., [2], [6], [12]), is to label the entire sequence based on the model that produces the maximum likelihood (MLC). MLC minimises the error under the assumption that the learnt distribution of the model represents the true distribution of the data. The disadvantage lies in this assumption as Hidden Markov Models (HMMs) are only approximations of the real process [4, 14]. We therefore, propose to turn this problem into a multidimensional classification problem, where the likelihood generated by every model represents one dimension. We refer to this approach as Likelihood Space Classification (LSC).

Likelihood Space Classification (LSC) has been inspired by [4] and [14]. [14] apply data space classification using mixture of Gaussians and LSC using linear discriminants for classifying texture and speech. [4] use GMMs for data space classification and LDA for LSC for classifying SAR images. To the best of our knowledge, LSC has never been investigated for multicue/muldimodal affect recognition and fusion. We obtain Likelihood Space Classification by using Support Vector Machines (SVM), a discriminative classifier widely explored in the field. Experimental results show that Likelihood Space Classification is superior to Maximum Likelihood Classification (MLC), and thus better suited to the task of audio-visual classification of spontaneous affective data. Furthermore, we expand the concept of LSC to address an open research question in the field which relates to the optimal way of fusing separate sets of cues in order to improve automatic affect recognition [5]. Our experiments show that projecting the likelihoods of HMMs trained on single cues and fusing them in likelihood space outperforms model-level fusion results of CHMMs projected onto the likelihood space.

2. Database

For the presented work we used the Sensitive Artificial Listener (SAL) Database [3]. It consists of spontaneous audio-visual data in the form of conversations that took place between a participant and an operator undertaking the role of an avatar with particular personalities. The recordings were made in a lab setting, using one camera, a uniform background and constant lighting conditions. SAL data has been annotated by 4 observers who provided continuous annotations with respect to valence and arousal dimensions. Although there are approximately 10 hours of footage available in the SAL database, A-V annotations have only been obtained for two female and two male subjects. For our experiments we used this portion, and based on the annotations provided, we automatically pre-segmented the audio-visual recordings into segments that contain positive or negative emotional expressions as described in detail in [10]. In total, we used 61 positive and 73 negative audio-visual segments, and 30,000 video frames.

Figure 1. Illustration of tracked points of (left) the face and (right) the shoulders.

3. Feature Extraction

Our audio feature vector consists of 15 features including the Mel-frequency Cepstrum Coefficients and prosody features (pitch, energy, RMSenergy), typically used for affect recognition from audio [17]. To capture the facial motion displayed during a spontaneous expression, the corners of the eyebrows (4 points), eyes (8 points), nose (3 points), mouth (4 points) and chin (1 point) are tracked using the Patras - Pantic particle filtering tracking scheme [11] (see Fig. 1). For each video segment containing n frames, the tracker results in a feature set with dimensions $n * 20 * 2$. The motion of the shoulders is captured by tracking 2 points on each shoulder and one stable point on the torso (see Fig. 1) by using the standard Auxiliary Particle Filtering [13]. The shoulder tracker results in a feature set with dimensions $n * 5 * 2$. For both shoulder and facial feature tracking, the points to be tracked were manually annotated in the first frame of an input video and tracked for the rest of the sequence.

4. Classification and Fusion

In human affective behaviour analysis, modality fusion refers to combining and integrating all incoming unimodal events into a single representation of the observed behaviour. Typically, multimodal data fusion is either done at the feature level in a maximum likelihood estimation manner or at the decision level when most of the joint statistical properties may have been lost. See [5, 17] for types and details of affective data fusion.

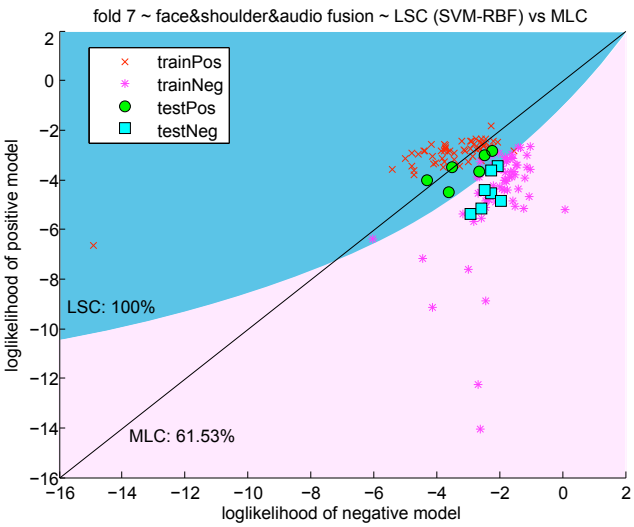

Figure 2. LSC decision surface (curved) vs. MLC decision surface (diagonal).

Model-level Fusion (MLF). In order to exploit the temporal correlation structure between the cues and modalities automatically via learning, we adopt model-level fusion based on Coupled Hidden Markov Models (CHMM). A CHMM is a series of parallel HMM chains coupled through cross-time and cross-chain conditional probabilities [9]. Therefore, CHMMs enable better modeling of intrinsic temporal correlations between multiple cues and modalities, and allow for true interactions between different feature sets corresponding to the same nonverbal display. In the HMM model, the probability of the next state of a sequence depends on the current state of the HMM. In the CHMM model the probability of the next state of a sequence depends on the current states of all HMMs.

Maximum Likelihood Classification (MLC). MLC finds the highest likelihood and determines the label of the input sequence based on the (C)HMM model that has produced it. MLC minimises the error under the assumption that the learnt distribution of the HMMs represents the true distribution of the data. The disadvantage lies in this assumption as HMMs are known to be only approximations of the real process [4, 14].

Likelihood Space Classification (LSC). Unlike MLC that simply determines the label based on the highest likelihood (C)HMM model, we propose to turn the classification problem into a multidimensional classification problem, where the likelihood generated by every (C)HMM model represents one dimension. Let us consider the two likelihoods which are the output of the (C)HMM models as points in a $2D$ space, with each dimension corresponding to the positive or negative class trained model. We refer to this space as the *likelihood space*. MLC bisects the $2D$ plane with the line $y = x$, which is used as a decision boundary for the classification (Fig. 2). As the learnt distribution by (C)HMMs is an approximation of the true distribution, we hypothesise that we can shift the line or even substitute it with a more complex function in order to achieve maximum separability. We refer to this approach as Likelihood Space Classification (LSC). All training samples are fed into the two trained (C)HMM models, and the likelihoods generated are projected onto the likelihood space where SVMs are trained by using the ground truth information. SVMs guarantee to find the optimal separating hyperplane in the feature space (mapped from the input space by a kernel function) given the defined parameters, minimizing the structural risk of the model (Fig. 2).

Likelihood Space Fusion (LSF). We further propose to turn the fusion problem into a multidimensional classification problem in the *likelihood space*, where the likelihood generated by every HMM model trained only for single cues, forms a feature vector $\mathbf{f} = \langle \theta_\mathbf{1} \dots \theta_\mathbf{c} \rangle$, where each $\theta_\mathbf{i}$ represents the pair of positive and negative likelihoods for cue i, i.e. $\theta_\mathbf{i} =< \theta_+, \theta_- >$. $\mathbf{f}$ then becomes a feature vector with dimensions $2 * c$, where c is the total number of fused cues.

5. Experiments and Results

Experimental Setup. The (C)HMM model size and state transition matrix for the face stream consists of four states, one for each temporal phase of neutral, onset, apex, and offset. The (C)HMMs used for the audio and shoulder sequences have three and two states respectively, while they are ergodic models. MLF is applied by using two-coupled and three-coupled HMMs, in order to accommodate two or three data streams. More details regarding the (C)HMM topology and parameter setting are discussed in [12].

In order to account for sequences of different length, we obtain the normalised log-likelihoods $\theta'_\mathbf{i}$ by dividing each pair of log-likelihoods generated by the (C)HMM models $\theta_\mathbf{i}$ on segment s by the number of frames in the respective segment $|s|$: $\theta'_\mathbf{i} = \frac{\theta_\mathbf{i}}{|s|}$ (similarly to [7]).

Since different combinations of cues produce a different distribution in the likelihood space, we tailor the learner parameters to specifically fit each case. 10-fold cross-validation has been used in all experiments, and classification accuracy, computed as the mean accuracy

Table 1. Experimental results for each combination of cues for MLC and LSC.

	F	S	A	FS	SA	FA	FSA
MLC	73.13%	73.88%	61.19%	78.36%	68.66%	70.90%	79.10%
LIN-LSC	90.27%	80.60%	69.01%	85.11%	75.38%	83.63%	87.20%
RBF-LSC	**91.76%**	**81.43%**	**72.97%**	**89.56%**	**80.60%**	**87.25%**	**90.98%**

of the 10 repetitions, is used as the performance measure. We perform subject-dependent recognition since the annotated part of SAL database contains data from 4 subjects only.

Experiment 1. In order to model the complex distribution of our data, we apply LSC utilising SVM with a linear and a Radial Basis Function (RBF) kernel (Table 1)[1]. The linear SVM kernel results show that the face cues gain the most important increase (more than 17%), attaining a classification accuracy of over 90%. The RBF kernel results in further improvement, showing that a more complex decision surface better separates the likelihood data (Fig. 2). Single facial expression cues and fusion of face&shoulder&audio cues provide a classification accuracy of over 90%, slightly higher than the fusion of face&shoulder cues (89.56%).

Experiment 2. As a significant increase in the classification accuracy has been achieved by adjusting the separation surface of the positive and negative 2D likelihood points, we perform LSF to evaluate how fusing the likelihoods of single-cue HMMs can compete with MLF obtained by CHMMs. Results for both the linear and the RBF kernel are presented in Table 2. In all cases and for both kernels, LSF improves the performance. The results for the RBF-LSF are similarly improved compared to model-level RBF-LSC. Many cue combinations achieve over 90% accuracy, while fusion of all cues reaches a classification accuracy of 94%. These results show that, for the data set at hand, LSF outperforms model-level-fused CHMMs for audio-visual classification of spontaneous affective data. The standard deviation of the number of incorrectly classified sequences per fold ($stdev$) is presented in Table 3. We observe that the most robust classifier appears to be the RBF-LSF for the fusion of face&shoulder&audio cues. The value of $stdev$ is generally $stdev \approx 1$, indicating a good level of consistency over the folds.

Analysis. From Table 1 (left half), we denote the superiority of the cue-specific RBF functions against the other approaches. Classification with RBF-LSF achieves results as high as 94.01% for the fusion of all cues (right half), whereas separation provided by MLC alone provides results of approximately 79%. Overall, for the task of positive vs. negative affect recognition from spontaneous data, the facial expression cues provide the best single cue recognition, followed by the shoulder and then the audio cues. The visual cues thus appear to be more significant than audio cues for automatic recognition of spontaneous affect in the valence dimension. The initial classification accuracy obtained with facial expression cues decreases slightly when fused with any other single cue (different cues possibly provide conflicting classification results for some sequences), while other single cue accuracies are greatly increased when fused with the facial expression cues. This does not come as a surprise since the SAL database contains subjects smiling while being ironic or people smiling while (non-)verbally expressing anger etc. Finally, the fusion of all cues and modalities provides us with the best classification results.

[1]The column headings of the tables are the initial letters of the cue(s) used in each experiment.

Table 2. Model-level fusion (with Likelihood Space Classification) vs. Likelihood Space Fusion.

SVM	FS	SA	FA	FSA
LIN-LSC	85.11%	75.38%	83.63%	87.20%
LIN-LSF	92.64%	**84.89%**	88.79%	91.09%
RBF-LSC	89.56%	80.60%	87.25%	90.98%
RBF-LSF	**93.41%**	84.18%	**91.70%**	**94.01%**

Table 3. Standard deviation ($stdev$) of the number of incorrectly classified sequences for each fusion method.

$stdev$	FS	SA	FA	FSA
RBF-LSC	0.843	0.966	0.948	1.032
RBF-LSF	0.994	0.875	1.197	0.823

6. Conclusion

This paper presented the first approach to focus on automatic classification of spontaneous affect from facial, shoulder and audio cues in the valence dimension. We hypothesised that for audio-visual classification of spontaneous affective data, when using dynamic gen-

erative models like (C)HMMs, MLC may not always be able to represent the model assumption accurately. We demonstrated with various experiments that projecting the results of generative classifiers onto likelihood space and then applying classification using discriminative classifiers such as SVMs has better robustness with regard to model specification.

Experimental results show that LSC is better suited to the task of audio-visual classification of spontaneous affective data than MLC, as it outperforms MLC with both linear and RBF kernels. Moreover, visual cues appear to be more significant than audio cues for automatic classification of spontaneous affect in the valence dimension. Finally, we introduced LSF and showed that it outperforms standard model-level fusion via CHMM combined with LSC, attaining a classification accuracy of 94.01% and improving all classification results. However, it should be noted that typically dynamic classifiers (e.g., (C)HMMs) are harder to train due to their complexity and number of parameters they need to learn [2]. In general, it is known that dynamic classifiers require more training samples compared to static classifiers: increasing the dimensionality does not seem to affect static classification, however, it visibly impedes the dynamic classification [12].

Although our results for the data set at hand appear to be robust, it remains an open issue whether LSF would still outperform model-level fusion with LSC when significantly higher number of data are available.

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). The work of Hatice Gunes has been funded by the European Communitys 7th Framework Programme [FP7/2007-2013] under the grant agreement no 211486 (SEMAINE).

References

[1] G. Caridakis, K. Karpouzis, and S. Kollias. User and context adaptive neural networks for emotion recognition. *Neurocomput.*, 71(13-15):2553–2562, 2008.

[2] I. Cohen, N. Sebe, L. Chen, A. Garg, and T. S. Huang. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and Image Understanding*, 91:160– 187, 2003.

[3] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, L. Lowry, M. McRorie, L. Jean-Claude Martin, J.-C. Devillers, A. Abrilian, S. Batliner, A. Noam, and K. Karpouzis. The humaine database: addressing the needs of the affective computing community. In *Proc. of the Second Int. Conf. on Affective Computing and Intelligent Interaction*, pages 488–500, 2007.

[4] R. Duan, W. Jiang, and H. Man. Robust adjusted likelihood function for image analysis. In *Proc. of IEEE Applied Imagery and Pattern Recognition Workshop*, volume 5237, pages 29 – 29, 2006.

[5] H. Gunes and M. Pantic. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions*, 1(1):68–99, 2010.

[6] H. Gunes and M. Piccardi. Automatic temporal segment detection and affect recognition from face and body display. *IEEE Tran. on Systems, Man, and Cybernetics-Part B*, 39(1):64–84, 2009.

[7] V.-B. Ly, S. Garcia-Salicetti, and B. Dorizzi. Fusion of HMM's likelihood and viterbi path for on-line signature verification. In *ECCV Workshop BioAW*, volume 3087 of *Lecture Notes in Computer Science*, pages 318–331. Springer, 2004.

[8] A. Mehrabian and J. Russell. *An Approach to Environmental Psychology*. Cambridge, New York, 1974.

[9] K. P. Murphy. The bayes net toolbox for matlab. *Computing Science and Statistics*, 33, 2001.

[10] M. Nicolaou, H. Gunes, and M. Pantic. Automatic segmentation of spontaneous data using dimensional labels from multiple coders. In *Proc. of Int. Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, in association with the Int. Conf. on Language Resources and Evaluation*, May 2010.

[11] I. Patras and M. Pantic. Particle filtering with factorized likelihoods for tracking facial features. In *International conference on automatic face and gesture recognition*, pages 97–104, 2004.

[12] S. Petridis, H. Gunes, S. Kaltwang, and M. Pantic. Static vs. dynamic modeling of human nonverbal behavior from multiple cues and modalities. In *Proc. of ACM Int. Conf. on Multimodal Interfaces*, pages 23–30, 2009.

[13] M. K. Pitt and N. Shephard. Filtering via simulation: auxiliary particle filters. *J. Am. Statistical Association*, 94(446):590–616, 1999.

[14] B. Raj and R. Singh. Classification in likelihood spaces. *Technometrics*, 46(3):318–329, 2004.

[15] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 1980.

[16] Wollmer, M. and Eyben, F. and Reiter, S. and Schuller, B. and Cox, C. and Douglas-Cowie, E. and Cowie, R. Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In *Proc. of 9th Interspeech Conf.*, pages 597–600, 2008.

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