Probabilistic Combination of Multiple Modalities to Detect Interest

Ashish Kapoor[†] Rosalind W. Picard[†] Yuri Ivanov[‡]

[†]Massachusetts Institute of Technology
Cambridge, MA 02139

[‡]Honda Research Institute US
Boston, MA 02139

Abstract

This paper describes a new approach to combine multiple modalities and applies it to the problem of affect recognition. The problem is posed as a combination of classifiers in a probabilistic framework that naturally explains the concepts of experts and critics. Each channel of data has an expert associated that generates the beliefs about the correct class. Probabilistic models of error and the critics, which predict the performance of the expert on the current input, are used to combine the expert's beliefs about the correct class. The method is applied to detect the affective state of interest using information from the face, postures and task the subjects are performing. The classification using multiple modalities achieves a recognition accuracy of 67.8%, outperforming the classification using individual modalities. Further, the proposed combination scheme achieves the greatest reduction in error when compared with other classifier combination methods.

1 Introduction

Many researchers have demonstrated that decisions from an ensemble of classifiers can be combined to improve classification accuracy. Multi-sensor classification is a problem that can be addressed by using classifier combination schemes. An ensemble of classifiers can be generated by training a classifier for each modality and the final classification can be performed using classifier combination methods. This paper describes a framework that can perform multi-sensor classification. The framework naturally gives rise to the concepts of experts and critics. Each expert performs classification using one channel of data, and for each expert there is a critic that predicts the performance of the expert. The predictions of the experts are combined with the evaluations of the critics for the final decision.

We demonstrate the multi-sensor classification scheme on the task of detecting the affective state of interest in children trying to solve a puzzle. The sensory information from the face, the postures and the state of the puzzle are combined in a probabilistic framework and we demonstrate that this method achieves a much better recognition accuracy than classification based on individual channels. Further, the critic-driven averaging [9], which is a special case of the proposed framework, outperforms all the other classifier combination methods applied to this problem.

2 Previous Work

There are many ensemble methods, including Boosting [12] and Bagging [1], that generate an ensemble of classifiers by choosing different samples from the training set. These methods require a common set of training data, which is a set of joint vectors formed by stacking the features extracted from all the modalities into one big vector. Often in multi-sensor fusion problems the training data has missing channels and labels, thus, most of the data cannot be used to form a common set of training data. Similarly, most of the data remains unused in the "feature-level fusion," where a single classifier is trained on joint features.

There has been other work on combining classifiers that avoid these problems. Kittler et al. [8] have described a common framework for combining classifiers and provide theoretical justification for using simple operators such as majority vote, sum, product, maximum and minimum. Hong and Jain [5] have used a similar framework to fuse multiple modalities for personal identification. The problem with these fixed rules is that, it is difficult to predict which rule would perform best. Miller and Yan [9] have provided a framework for critic driven ensemble classification. Their ensemble scheme consists of base level classifiers called experts, and for each of these experts there is a second level classifier called a critic that predicts how well an expert is going to perform on the current input. The decisions from the experts are combined using the evaluation by the critics. Like Miller and Yan, we use critic-based combination method, but tackle the problem from a Bayesian perspective and show that the critic driven ensemble classification is a special case of our scheme.

Also, there have been advances in machine recognition of human emotion [10, 11] and most of the work so far has relied on a single modality. Exceptions include Huang et al. [6], who have used audio and video to recognize six different affective states of happiness, sadness, anger, fear, surprise and disgust. The work reported here, extends the work by *Mota* [10] where she demonstrated a system that uses just the posture information to detect interest. To our knowledge, our work is the first of its kind where an attempt is made to use multiple modalities for the purpose of detecting interest.

The next section describes the scheme for combining multiple modalities. Followed by that, we describe the overall framework and the different modalities used for the purpose of affect recognition. After that, we conclude with experimental evaluations and future work.

3 Combining Multiple Modalities

Our approach to sensor fusion is based on viewing the output of individual classifiers as a random variable, denoted by $\tilde{\omega}$. If m is an indicator variable that corresponds to the modality that is used for classification, then as shown in [7], the probability of the true class label, ω, for a given observation $\mathbf{x}$ can be written as:

$$P(\omega|\mathbf{x}) = \sum_{m=1}^{n} P(m|\mathbf{x}) \sum_{\tilde{\omega}} P(\tilde{\omega}|\mathbf{x}, m) P(\omega|\tilde{\omega}, \mathbf{x}, m) \qquad (1)$$

Equation 1 can be used as a framework to combine the base level classifiers that correspond to individual modalities. The term $P(\tilde{\omega}|\mathbf{x}, m)$ corresponds to the base level classifiers (experts). $P(m|\mathbf{x})$ corresponds to the classifiers that are critics, which evaluate how well each expert can perform on the input. The term $P(\omega|\tilde{\omega}, \mathbf{x}, m)$ models the error each expert makes on the current input. In the work described here, we approximate the error term $P(\omega|\tilde{\omega}, \mathbf{x}, m)$ as $P(\omega|\tilde{\omega}, m)$. Hence:

$$P(\omega|\mathbf{x}) \approx \sum_{m=1}^{n} P(m|\mathbf{x}) \sum_{\tilde{\omega}} P(\tilde{\omega}|\mathbf{x}, m) P(\omega|\tilde{\omega}, m) \qquad (2)$$

Given some training data, we can easily train the experts and the critics. The term $P(\omega|\tilde{\omega}, m)$ is an empirical distribution and can be obtained from the confusion matrix generated by testing the experts on the training data. Also, note that if we assume that $P(\omega|\tilde{\omega}, m) = 1$ for all $\omega = \tilde{\omega}$, we derive the critic-based averaging (Miller and Yan [9]), proving that it is a special case of the framework derived here.

One of the main advantages of this framework is that the individual classifiers can be trained separately; thus extending a multimodal system to incorporate new modalities is easy. Further, training data corresponding to each modality can be separately collected rather than collecting training data that consists of all the modalities synchronized together. On the other hand this approach cannot use any relationship between the different channels, which could have been used for classification. These kind of relationships can be useful in cases like audio-video fusion. In our application, we do not face this problem, as the channels used are mostly conditionally independent given the class label.

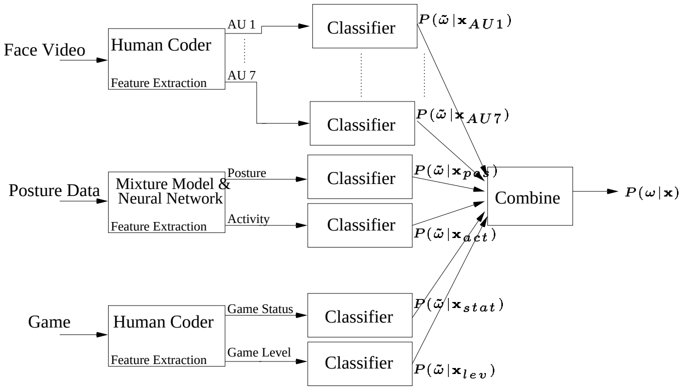
Face Video
Human Coder
Feature Extraction
AU 1
AU 7
Classifier
Classifier
$P(\tilde{\omega}|\mathbf{x}_{AU1})$
$P(\tilde{\omega}|\mathbf{x}_{AU7})$
Posture Data
Mixture Model & Neural Network
Feature Extraction
Posture
Activity
Classifier
Classifier
$P(\tilde{\omega}|\mathbf{x}_{pos})$
$P(\tilde{\omega}|\mathbf{x}_{act})$
Combine
$P(\omega|\mathbf{x})$
Game
Human Coder
Feature Extraction
Game Status
Game Level
Classifier
Classifier
$P(\tilde{\omega}|\mathbf{x}_{stat})$
$P(\tilde{\omega}|\mathbf{x}_{lev})$

Figure 1: The overall architecture

4 Detecting Interest

We applied the framework described in the previous section to the problem of machine recognition of affect using multiple modalities. We look at the problem of detecting the affective states of high-interest, low-interest and “refreshing” in a child who is solving a puzzle. We define “refreshing” as a state associated with a short break in the task to restore concentration again. There are many applications of this system like intelligent tutoring systems, adaptive interfaces etc. The scenario we focus on has a child solving a puzzle and a machine trying to infer the affective state using the face, the postures and the information from the puzzle.

The overall system architecture is shown in figure 1. The raw data from the camera, the posture sensor and the game being played is first analyzed to extract relevant features. There are classifiers (experts) that correspond to each of these extracted features. Based on the extracted features, all of these experts predict the affective state ($P(\tilde{\omega}|\mathbf{x_i})$ for all i). These predictions are combined, as described in section 3, using critics and error estimates. The details of the modalities, extracted features and experts are given below.

4.1 Modality 1: Facial Actions

The facial actions can often tell us about the affective and the cognitive state of the person [4]. The features extracted from the videos of the face correspond to the upper facial action units described in the Facial Action Coding System (FACS). The Facial Action Coding System (FACS) developed by Ekman and Friesen [3] is a method of measuring facial activity in terms of the facial muscle movements. The upper action units (Action Units 1,2,4,5,6 and 7) correspond to the muscle movements around the eyes, eyebrows and upper cheek. Our feature extraction is manual, where a certified FACS coder looked at the videos of the face and coded them for the presence of facial action units and their combinations. Specifically, from each video of the face database, the feature extraction produces six sequences, which correspond to the presence or absence of an action unit in each video frame. Although, in this work the FACS coder manually labeled the video data, this can be automated using computer vision systems that can detect facial actions automatically from frontal video of the face [13].

4.2 Modality 2: Postures

A sensor chair that uses an array of force sensitive resistors is used to sense postures in our system. *Mota* [10] has shown that the postures can be used to recognize the affective state of interest. Our work is an extension where we use the same posture sensing system. The sensor chair has two 0.10 mm thick sensor sheets, which are arrays of 42-by-48 sensing units. Each unit outputs an 8-bit pressure reading. One of the sheets is placed on the backrest and one on the seat. The pressure distribution map (2 of 42x48 points) sensed at a sampling frequency of 50Hz is used to infer information about the posture. *Mota* [10] has used a mixture of Gaussians to represent the pressure patterns. A Neural Network further classifies these representations for 8 different postures and 3 different levels of activity. We use these labels for postures and the level of activity as the two features from this modality.

4.3 Modality 3: Game State Information

There are two features extracted from the game. They correspond to the state of the game and the level of difficulty. The state of the game indicates if a new game has started or ended, or whether the child asked for a hint, or whether the solution provided was correct etc. In this work, all these features were manually extracted. Specifically, given the sequence of screen captures, the feature extraction module produces two sequences corresponding to game state and the level being played, for every sample.

5 Implementation & Results

Given a sequence of all the 10 extracted features (6 from the face, 2 from the posture and 2 from the game), we can use the classifier combination method described in section 3 to infer information about the affective state. In our implementation, we use Hidden Markov Models (HMMs) as experts and critics. The training phase consists of first training the experts. Once the experts are trained, they are tested on the training data itself. The critics are trained on the success and failures of the experts on the training data. The confusion matrix is used to recover the error model ($P(\omega|\tilde{\omega}, m)$). During the test phase, each expert generates its beliefs about the class the sequence belongs to, which are then combined using equation 2.

To evaluate our method on real-life situations, we conducted a study to collect a multimodal database. The subjects were 9 children who were asked to play a game called the *fripple place* [2]. The game has a number of puzzles that require mathematical reasoning. While each subject worked on these puzzles for about 20 minutes, three channels of data (face, posture and game information) were recorded. To find out the ground truth for the affective state, several teachers looked at the database we recorded, and annotated the data segments for the affective states of high interest, low interest, refreshing, bored, neutral and the ”other” category. The database we chose only consists of data-segments on which the teachers agreed on the label of affective states. The details of annotation can be found in [10]. The database included all the 8 children (all three channels were available for 4 children and only posture and game for other 4) and consisted of 98 samples of high-interest, 94 samples of low-interest and 70 samples of refreshing. Each of the samples is a maximum of 8 secs long with observations recorded at 8 samples per second.

We also evaluate the performance of other classifier combination schemes. Table 1 describes how these combination schemes are used to combine experts. Also, we evaluate critic-based-averaging [9] and performance based combination. To implement critic-based-averaging we simply set $P(\omega|\tilde{\omega}, m) = 1$ for all $\omega = \tilde{\omega}$ in equation 2. The performance based combination is again a special case of the framework where, $P(m|\mathbf{x})$ is uniform for all m. Which means, that all the critics are turned off and the modalities are weighed just based upon their performance on the training data. Once the modalities are combined, for all the combination methods the final class label is chosen such that it maximizes the combined posterior. That is, $\omega^* = \arg\max_\omega P(\omega|\mathbf{x})$.

60% of the data is randomly selected as the training data, while the rest is treated as the testing data. Each of the ex-

Table 1: Classifier Combination Methods

Combination Rule	**Criteria**
Sum	$\forall i, P(\omega = i|\mathbf{x}) \propto \sum_{m=1}^{n} P(\tilde{\omega} = i|\mathbf{x}, m)$
Product	$\forall i, P(\omega = i|\mathbf{x}) \propto \prod_{m=1}^{n} P(\tilde{\omega} = i|\mathbf{x}, m)$
Max	$\forall i, P(\omega = i|\mathbf{x}) \propto \max_m P(\tilde{\omega} = i|\mathbf{x}, m)$
Min	$\forall i, P(\omega = i|\mathbf{x}) \propto \min_m P(\tilde{\omega} = i|\mathbf{x}, m)$
Majority Voting	$\forall i, P(\omega = i|\mathbf{x}) \propto$ $\{ \begin{matrix} 1 & \text{if } i = \text{vote}_m(\arg\max_{\tilde{\omega}} P(\tilde{\omega}|\mathbf{x}, m)) \\ 0 & \text{otherwise} \end{matrix} \}$

Table 2: Averaged Results: Individual Modalities

Modality	Recognition Rate
Facial Action Unit 1	49.7%
Facial Action Unit 2	48.6%
Facial Action Unit 4	32.8%
Facial Action Unit 5	38.1%
Facial Action Unit 6	42.4%
Facial Action Unit 7	36.0%
Postures	55.1%
Activity on Chair	60.1%
Game Status	33.0%
Game: Difficulty Level	25.4%

Table 3: Averaged Results: Combining Modalities

Combining Scheme	Average Recognition Rate	Average Reduction in Error
Product	62.6%	0.5%
Addition	60.7%	-4.9%
Vote	55.9%	-16.8%
Max	54.3%	-21.5%
Min	60.1%	-6.2%
Performance-based Weighing	65.1%	7.1%
Critic-based Averaging	65.9%	9.3%
Full Method	67.8%	14.1%

perts and the critics are trained and all of the combination methods are tested using these experts and the critics. The process of randomly splitting the database, training and testing is performed 50 times and the average recognition rates are reported as results. Table 2 & 3 show the average recognition results, for 50 runs of test, using the individual channels and by combining modalities respectively. From Table 2 it can be seen that the posture activity channel is most discriminating with 60.1% recognition accuracy. For the classifier combination methods, we also measure the reduction in error for all the combination methods. The reduction in error for a combination method in a test run is defined as the percentage reduction in error with respect to the error made by the classifier trained on the posture activity, which was the best discriminating single modality overall. Table 3 shows that our method performs best and outperforms other classifier combination methods with an average accuracy of 67.8% and 14.1% average reduction in error. The Critic-based Averaging achieves the next best average recognition rate of 65.9% and 9.3% average reduction in error. We hypothesize that the perfromance can be improved by exactly modeling the errors rather than approximating $P(\omega|\tilde{\omega}, m, \mathbf{x})$ as $P(\omega|\tilde{\omega}, m)$. Although, modeling these errors increase the complexity of the system, it should be able to significantly improve the performance.

6 Conclusion and Future Work

We described a framework to build multimodal classification systems that uses experts, critics & error models to combine classifiers. One of the advantages of this framework is that the classifiers can be separately trained, thus extending a system to incorporate new modalities is easy. The framework is used to detect the affective state of interest. We achieve an average recognition accuracy of 67.8%, which is more than the accuracy obtained using any single modality. Further, the approach outperforms other standard classifier combination methods. Future work includes finding schemes that would also model the relationship accross modalities. There are many applications of this framework where information from many sources needs to be integrated to infer about some quantity of interest.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 0087768.

References

[1] L. Breiman. Bagging predictors. *Machine Learning*, 26(2), 1996.

[2] Edmark. Fripple place. http://www.riverdeep.net/edconnect/softwareactivities /critical_thinking/fripple_place.jhtml.

[3] P. Ekman and W. V. Friesen. *The Facial Action Coding System: A Technique for Measurement of Facial Movement*. Consulting Psychologists Press, San Francisco, CA, 1978.

[4] J. Hager and P Ekman. Essential behavioral science of the face and gesture that computer scientists need to know. In *International Workshop on Automatic Face and Gesture Recognition*, June 1996.

[5] L. Hong and A. K. Jain. Integrating faces and fingerprints for personal identification. *Pattern Analysis and Machine Intelligence*, 20(12), 1998.

[6] Thomas S. Huang, Lawrence S. Chen, and Hai Tao. Bimodal emotion recognition by man and machine. In *Proceedings of ATR Workshop on Virtual Communication Environments*, April 1998.

[7] Y. Ivanov, T. Serre, and J. Bouvrie. Error weighted classifier combination for multi-modal human identification. In *Submission*, 2004.

[8] Josef Kittler, Mohamad Hatef, Robert P. W. Duin, and Jiri Matas. On combining classifiers. *Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[9] David J. Miller and Lian Yan. Critic-driven ensemble classification. *Signal Processing*, 47(10):2833–2844, 1999.

[10] Selene Mota. Automated posture analysis for detecting learner's affective state. Master's thesis, MIT, 2002.

[11] Yuan Qi, Carson Reynolds, and Rosalind W. Picard. The bayes point machine for computer user frustration detection via pressure mouse. In *Proceedings from the workshop on Perceptive User Interface*, 2001.

[12] R. Schapire. A brief introduction to boosting. In *Proceedings of International Conference on Algorithmic Learning Theory*, 1999.

[13] Y. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *Pattern Analysis and Machine Intelligence*, 23(2), February 2001.