Multimodal Affect Recognition in Learning Environments

Ashish Kapoor and Rosalind W. Picard
MIT Media Laboratory
Cambridge, MA 02139, USA
{kapoor, picard}@media.mit.edu

ABSTRACT

We propose a multi-sensor affect recognition system and evaluate it on the challenging task of classifying interest (or disinterest) in children trying to solve an educational puzzle on the computer. The multimodal sensory information from facial expressions and postural shifts of the learner is combined with information about the learner's activity on the computer. We propose a unified approach, based on a mixture of Gaussian Processes, for achieving sensor fusion under the problematic conditions of missing channels and noisy labels. This approach generates separate class labels corresponding to each individual modality. The final classification is based upon a hidden random variable, which probabilistically combines the sensors. The multimodal Gaussian Process approach achieves accuracy of over 86%, significantly outperforming classification using the individual modalities, and several other combination schemes.

Categories and Subject Descriptors

I.5 [**Computing Methodologies**]: Pattern Recognition; I.4.9 [**Image Processing and Computer Vision**]: Applications; J.4 [**Computer Applications**]: Social and Behavioral Sciences

General Terms

Algorithms, Design, Human Factors, Performance

1. INTRODUCTION

We present a framework to automatically extract, process and model sequences of natural occurring non-verbal behavior for recognizing affective states that occur during natural learning situations. The framework will be a component in computerized learning companions [1, 6] that could provide effective personalized assistance to children engaged in learning explorations and will also help in developing theoretical understanding of human behavior in learning situations.

MM'05, November 6–11, 2005, Singapore.

This work tackles a number of challenging issues. While most of the prior work on emotion recognition has focused on posed emotions by actors, our emphasis has been on naturally occurring non-verbal behaviors as it is crucial that the system is capable of dealing with the unconstrained nature of real data. Second, despite the advances in face analysis and gesture recognition, the real-time sensing of non-verbal behaviors is still a challenging problem. In this work we demonstrate a multi-modal system that can automatically extract non-verbal behaviors and features from face and postures, which can be used to detect affective states. Third, training a pattern classification system needs labeled data. Unlike the case of posed emotions, in natural data getting the ground truth of labels is a challenging task. We discuss how we can label the data reliably for different affective states. Moreover, we note that there is always an uncertainty about the true labels of the data. There might be labeling noise; that is, some data points might have incorrect labels; thus, requiring a principled approach to handle any labeling noise in the data. Finally, we address pattern recognition in the multi-modal scenario, which has been addressed previously by using either data-level fusion or classifier combination schemes. In the former, a single classifier is trained on joint features; however, the sensors might often fail and result in missing or bad data, a frequent problem in many multimodal scenarios, resulting in a significant reduction in the performance of the pattern recognition system.

We demonstrate an affect recognition system that addresses all four challenges mentioned above. The goal of the system is to classify affective states related to interest in children trying to solve a puzzle on a computer. The system uses real-time tracking of facial features and behaviors and monitors postures to extract relevant non-verbal cues. The extracted sensory information from the face, the postures and the state of the puzzle are combined using a unified Bayesian approach based on a mixture of Gaussian Process classifiers where classification using each channel is learned via Expectation Propagation [11]. The decision about the affective state is made by combining the beliefs of each individual expert using another meta-level classification system. The approximate Bayesian inference in the model is very efficient as EP approximates the probability distribution over each expert's classification as a product of Gaussians and can be updated very quickly. The system is evaluated on natural data and it achieves an accuracy of over 86%, significantly outperforming classification using the individual modalities and several other combination schemes.

2. PREVIOUS WORK

A lot of research has been done to develop methods for inferring affective states. Many researchers have used static methods such as questionnaires, dialogue boxes, etc., which are easy to administer but have been criticized for being static and thus not able to recognize changes in affective states. A more dynamic and objective approach for sensing affect is via sensors such as cameras, microphones, wearable devices, etc. However, most of the work on affect recognition using the sensors focuses on deliberately expressed emotions (happy /sad /angry etc.) by actors, and not on those that arise in natural situations such as classroom learning. In the context of learning there have been very few approaches for the purpose of affect recognition. Notable among them is Conati's [2] work on probabilistic assessment of affect in educational games. Also Mota and Picard [12] have described a system that uses dynamic posture information to classify different levels of interest in learning environments, which we significantly extend to the multimodal scenario.

Despite the advances in machine recognition of human emotion, much of the work on machine recognition of human emotion has relied on a single modality. Exceptions include the work of Picard et al.[15], which achieved 81% classification accuracy of eight emotional states of an individual over many days of data, based on four physiological signals, and several efforts that have combined audio of the voice with video of the face, e.g. Huang et al. [3], who combined these channels to recognize six different affective states. Pantic and Rothkrantz [14] provide a survey of other audio-video combination efforts and an overview of issues in building a multimodal affect recognition system.

A lot of researchers have looked into the general problem of combining information from multiple channels. A common approach is "feature-level fusion", where a single classifier is trained on joint features, formed by stacking the features extracted from all the modalities into one big vector. Often in affect sensing scenarios, the sensors or feature extraction might fail, resulting in data with missing channels and reducing the performance of this approach. One alternate solution is to use decision-level fusion. Kittler et al. [9] have described a common framework for combining classifiers and provided theoretical justification for using operators such as vote, sum, product, maximum and minimum. One problem with these fixed rules is that it is difficult to predict which rule would perform best. There are methods, such as layered HMMs [13], which perform decision fusion and sensor selection depending upon utility functions and stacked classifiers. One main disadvantage of using stacked based classification is that these methods require a large amount of training data. Alternatively, there are mixture-of-experts [4] and critic-driven approaches [10] where base-level experts are combined using critics that predict how well an expert is going to perform on the current input. In similar spirit Toyama and Horvitz [16] demonstrate a head tracking system that uses contextual features as reliability indicators to select different algorithms. In an earlier work [8] we have proposed an expert-critic system based on HMMs to combine multiple modalities. In this paper we show an alternative fusion strategy within the Bayesian framework which significantly beats the earlier method.

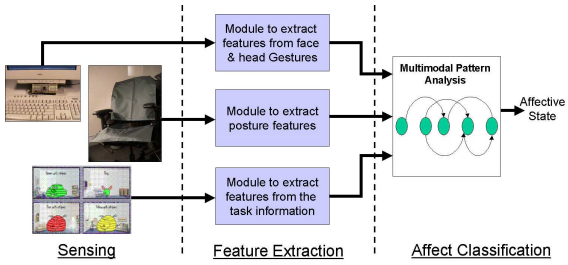


Figure 1: The overall architecture

3. THE PROPOSED FRAMEWORK

Figure 1 describes the architecture of the proposed system. The non-verbal behaviors are sensed through a camera and a pressure sensing chair. The camera is equipped with Infrared (IR) LEDs for structured lighting that help in real-time tracking of pupils and extracting other features from the face. Similarly the data sensed through the chair is used to extract information about the postures. Features are extracted from the activity that the subject is doing on the computer as well, which are then sent to a multimodal pattern analyzer that combines all the information to predict the current affective state. In this paper we focus on a scenario where children try to solve puzzles on a computer.

3.1 Facial Features & Head Gestures

The feature extraction module for face and head gestures is shown in figure 2. We use an in-house built version of the IBM Blue Eyes Camera that tracks pupils unobtrusively using two sets of IR LEDs. One set of LEDs is on the optical axis and produces the red eye effect. The two sets of LEDs are switched on and off to generate two interlaced images for a single frame.The image where the on-axis LEDs are on has white pupils whereas the image where the off-axis LEDs are on has black pupils. These two images are subtracted to get a difference image, which is used to track the pupils. The pupils are detected and tracked using the difference image, which is noisy due to the motion artifacts and other specularities. We have elsewhere described [7] an algorithm to track pupils reliably using the noisy difference image. Once tracked, the pupil positions are passed to an HMM based head-nod and head-shake detection system,which provides the likelihoods of head-nods and head-shakes. Similarly, we have also trained an HMM that uses the radii of the visible pupil as inputs to produce the likelihoods of blinks. Further, we use the system described in [7] to recover shape information of eyes and the eyebrows. Given pupil positions we can also localize the image around the mouth. Rather than extracting the shape of the mouth we extract two real numbers which correspond to two kinds of mouth activities: smiles and fidgets. We look at the sum of the absolute difference of pixels of the extracted mouth image in the current frame with the mouth images in the last 10 frames. A large difference in images should correspond to mouth movements, namely the fidgets. Besides a numerical score that corresponds to fidgets, the system also uses a support vector machine (SVM) to compute the probability of smiles. Specifically, an SVM was trained using natural examples of mouth images, to classify mouth images as smiling or not smiling. The localized mouth image in the current

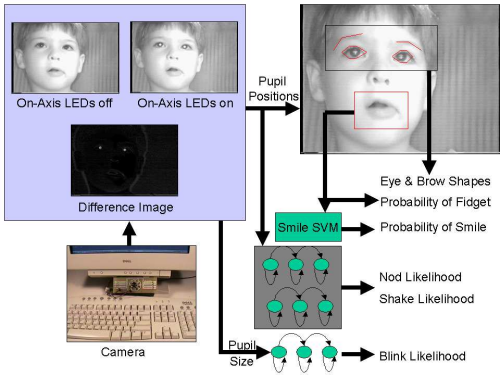


Figure 2: Module to extract facial features.

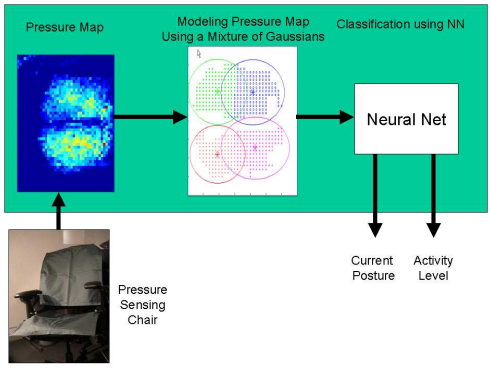


Figure 3: Module to extract posture features.

frame is used as an input to this SVM classifier and the resulting output is passed through a sigmoid to compute the probability of smile in the current frame. The system can extract features in real time at 27-29 frames per second on a 1.8 GhZ Pentium 4 machine. The system tracks well as long as the subject is in the reasonable range of the camera. The system can detect whenever it is unable to find eyes in its field of view, which might occasionally happen due to large head and body movements. Also, sometimes the camera can only see the upper part of the face and cannot extract lower facial features, which happens if the subject leans forward. Due to these problems we often have missing information from the face; thus, we need an affect recognition system that is robust to such tracking failures.

3.2 The Posture Sensing Chair

Postures are recognized using two matrices of pressure sensors made by Tekscan. One matrix is positioned on the seat-pan of a chair; the other is placed on the backrest. Each matrix is 0.10 millimeters thick and consists of a 42-by-48 array of sensing pressure units distributed over an area of 41 x 47 centimeters. A pressure unit is a variable resistor, and the normal force applied to its superficial area determines its resistance. This resistance is transformed to an 8-bit pressure reading, which can be interpreted as an 8-bit grayscale value and visualized as a grayscale image. Figure 3 shows the feature extraction strategy used for postures in [12]. First, the pressure maps sensed by the chair are pre-processed to remove noise and the structure of the map is modeled with a mixture of Gaussians. The parameters of the Gaussian mixture (means and variances) are used to feed a 3-layer feed-forward neural network that classifies the static set of postures (for example, sitting upright, leaning back, etc.) and activity level (low, medium and high) in real time at 8 frames per second, which are then used as posture features by the multimodal affect classification module.

4. RECOGNIZING AFFECTIVE STATES

Table 1 shows all the features that are extracted every one eighth of a second. We deliberately grouped them under "channels", separating for example the upper and lower face features because often the upper face features were present but not the lower. We form four different channels (table 1) and each channel corresponds to a group of features that can go missing simultaneously depending upon which sensor fails. Our strategy will be to fuse decisions from these channels, rather than individual features, thus, preserving some higher order statistics between different features.

Figure 4 shows the model we follow to solve the problem. The data $\mathbf{x}^p$ from P different channels generate soft-decisions y^p corresponding to each different channel. The variable λ determines the channel that decides the final decision t. Note, that λ is never observed and we have to marginalize over λ to compute the final decision. Intuitively, this can be thought of as weighting individual decisions y^p appropriately and combining them to infer the distribution over the class label t. The weights, which corresponds to a probability distribution over λ, depend upon the data point being classified and are determined using another meta-level decision system. While the results in this paper are obtained with λ selecting to one channel at a time, it is possible more generally to have it select multiple channels. The system described in this paper uses Gaussian Process (GP) classification to first infer the probability distribution over y^p for all the channels. The final decision is gated through λ whose probability distribution conditioned on the test point is determined using another multi-label GP classification system.

We follow a Bayesian paradigm and the aim is to compute the posterior probability of an affective label of a test point given all the training data and the model. The next subsection describes the acquisition of the annotated training data. Following that, we review GP classification using EP and then show how to extend the idea to a Mixture of GPs in order to handle multiple modalities in the same framework.

Table 1: Extracted features from different modalities which are grouped into channels.

Channel 1: Upper Face
Brow Shape
Eye Shape
likelihood of nod
likelihood of shake
likelihood of blink

Channel 2: Lower Face
Probability of Fidget
Probability of Smile

Channel 3: Posture
Current Posture
Level of Activity

Channel 4: Game
Level of Difficulty
State of the Game

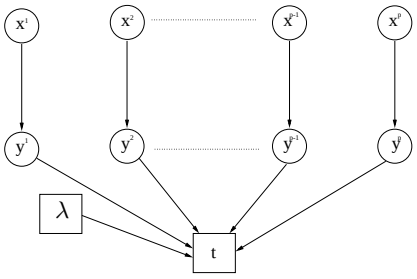


Figure 4: A mixture of GPs for p channels

4.1 Collecting the Database

Data collection for affective studies is a challenging task: The subject needs to be exposed to conditions that can elicit the emotional state in an authentic way; if we elicit affective states on demand, it is almost guaranteed not to bring out genuinely the required emotional state. Affective states associated with interest and boredom were elicited through an experiment with children aged 8 to 11 years, coming from relatively affluent areas of the state of Massachusetts in the USA. Each child was asked to solve a constraint satisfaction game called Fripples Place for approximately 20 minutes and the space where the experiment took place was a naturalistic setting allowing the subject to move freely. It was arranged with the sensor chair, one computer playing the game and recording the screen activity, having a monitor, standard mouse and keyboard, as well as two video-cameras: one capturing a side-view and one the frontal view; and finally, a Blue Eyes camera capturing the face. We made the cameras unobtrusive to encourage natural responses preserving as much as possible the original behavior. Given that we cannot directly observe the student's internal thoughts and emotions, nor can children in the age range of 8 and 11 years old reliably articulate their feelings, we chose to focus on labeling affective states by observers who are teachers by profession. We engaged in several iterations with teachers to ascertain a set of meaningful labels that could be reliably inferred from the data. The teachers were allowed to look at frontal video, side video, and screen activity, recognizing that people are not used to looking at chair pressure patterns. Eventually, we found that teachers could reliably label the states of high, medium and low interest, bored, and a state that we call "taking a break," which typically involved a forward-backward postural fidget and sometimes stretching. Working separately and without being aware of the final purpose of the coding task, teachers obtained an average overall agreement (Cohen's Kappa) of 78.6%. In this work, we did not use data classified as "bored" or "other" even though teachers identified them consistently. The bored state was dropped since teachers only classified very few episodes as bored, and this was not enough to develop separate training and test sets. The final database used to train and test the system included 8 different children with 61 samples of "high interest," 59 samples of "low interest" and 16 samples of "taking a break." Each of the samples is a maximum of 8 secs long with observations recorded at 8 samples per second. Only 50 samples had features present from all the four modalities, whereas the other 86 samples had the face channel missing.

4.2 Gaussian Process Classification

GP classification is related to kernel machines such as SVMs and has been well explored in the machine learning community. Under the Bayesian framework, given a set of labeled data points $\mathbf{X} = \{\mathbf{x}_1, .., \mathbf{x}_n\}$, with class labels $\mathbf{t} = \{t_1, .., t_n\}$ and an unlabeled point $\mathbf{x}^*$, we are interested in the distribution $p(t^*|\mathbf{X}, \mathbf{t}, \mathbf{x}^*)$. Here t^* is a random variable denoting the class label for the point $\mathbf{x}^*$. The idea behind GP classification is that the hard labels $\mathbf{t}$ depend upon hidden soft-labels $\mathbf{y} = \{y_1, ..., y_n\}$, which are assumed to be jointly Gaussian with the covariance between two outputs y_i and y_j specified using a kernel function applied to $\mathbf{x}_i$ and $\mathbf{x}_j$. Formally, $\{y_1, .., y_n\} \sim \mathcal{N}(0, \mathbf{K})$ where $\mathbf{K}$ is a n-by-n kernel matrix with $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. The observed labels $\mathbf{t}$ are assumed to be conditionally independent given the soft labels $\mathbf{y}$ and each t_i depends upon y_i through the conditional distribution:

$$p(t_i|y_i) = \Phi(\beta y_i \cdot t_i)$$

Here, $\Phi(z) = \int_{-\infty}^{z} \mathcal{N}(z; 0, 1)$ which provides a quadratic slack for labeling errors; thus, the model should be more robust to label noise. We are also looking at other Bayesian techniques to handle label noise and uncertainty, which we will report in a future publication.

Our task is to infer $p(t^*|D)$, where $D = \{\mathbf{X}, \mathbf{t}, \mathbf{x}^*\}$. We use Expectation Propagation (EP) to first approximate $P(\mathbf{y}, y^*|D)$ as a Gaussian and then use the approximate distribution $p(y^*|D) \approx \mathcal{N}(M^*, V^*)$ to classify the test point $\mathbf{x}^*$:

$$p(t^*|D) \propto \int_{y^*} p(t^*|y^*)\mathcal{N}(M^*, V^*) \tag{1}$$

One of the useful byproducts of EP is the Gaussian approximations of the likelihoods $p(t_i|y_i)$:

$$p(t_i|y_i) \approx \tilde{t}_i = s_i \exp(-\frac{1}{2v_i}(y_i \cdot t_i - m_i)^2) \tag{2}$$

Conceptually, we can think of EP starting with the GP prior $\mathcal{N}(0, \mathbf{K})$ over the hidden soft labels $(\mathbf{y}, y^*)$ and incorporating all the approximate terms $\tilde{t}_i$ to approximate the posterior $p(\mathbf{y}, y^*|D) = \mathcal{N}(\mathbf{M}, \mathbf{V})$ as a Gaussian. For details readers are encouraged to look at [11].

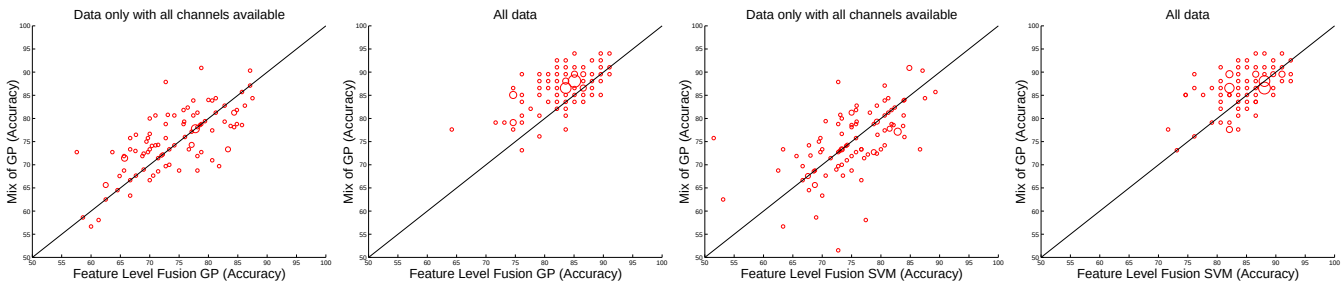
4.3 Mixture of GPs for Sensor Fusion

Given n data points $\bar{\mathbf{x}}_1, .., \bar{\mathbf{x}}_n$, obtained from P different sensors, our approach follows a mixture of GP model described in figure 4. Let every i^{th} data point be represented as $\bar{\mathbf{x}}_i = \{\mathbf{x}_i^{(1)}, .., \mathbf{x}_i^{(P)}\}$, and the soft labels as $\bar{\mathbf{y}}_i = \{y_i^{(1)}, .., y_i^{(P)}\}$. Given $\lambda_i \in \{1, .., P\}$, the random variable that determines the channel for each point's classification, the classification likelihood can be written as:

$$P(t_i|\bar{\mathbf{y}}_i, \lambda_i = j) = P(t_i|y_i^{(j)}) = \Phi(\beta t_i \cdot y_i^{(j)})$$

Given a test point $\bar{\mathbf{x}}^*$, let $\bar{\mathbf{X}} = \{\bar{\mathbf{x}}_1, .., \bar{\mathbf{x}}_n, \bar{\mathbf{x}}^*\}$ denote all the training and the test points. Further, let $\bar{\mathbf{Y}} = \{\mathbf{y}^{(1)}, .., \mathbf{y}^{(P)}\}$, denote the hidden soft labels corresponding to each channel of all the data including the test point. Let, $Q(\bar{\mathbf{Y}}) = \prod_{p=1}^{P} Q(\mathbf{y}^{(p)})$ and $Q(\mathbf{\Lambda}) = \prod_{i=1}^{n} Q(\lambda_i)$, denote the approximate posterior over the hidden variables $\bar{\mathbf{Y}}$ and $\mathbf{\Lambda}$, where $\mathbf{\Lambda} = \{\lambda_1, .., \lambda_n\}$ are the switches corresponding only to the n labeled data points. Let $p(\bar{\mathbf{Y}})$ and $p(\mathbf{\Lambda})$ be the priors with $p(\bar{\mathbf{Y}}) = \prod_{p=1}^{P} p(\mathbf{y}^{(p)})$, the product of GP priors and $p(\mathbf{\Lambda})$ uniform. Our algorithm aims to compute good approximations $Q(\bar{\mathbf{Y}})$ and $Q(\mathbf{\Lambda})$ to the real posteriors by iteratively optimizing the variational bound:

$$F = \int_{\bar{\mathbf{Y}}, \mathbf{\Lambda}} Q(\bar{\mathbf{Y}})Q(\mathbf{\Lambda}) \log(\frac{p(\bar{\mathbf{Y}})p(\mathbf{\Lambda})p(\mathbf{t}|\bar{\mathbf{X}}, \bar{\mathbf{Y}}, \mathbf{\Lambda})}{Q(\bar{\mathbf{Y}})Q(\mathbf{\Lambda})}) \tag{3}$$



(a) (b) (c) (d)

Figure 5: Plots comparing accuracy of Mixture GP with GP (a)-(b) and SVM (c)-(d) where the latter methods use feature fusion. In (a) and (c) only the subset of data where all modes are intact is used. In (b) and (d) all the data is used, as described in the text. There are 100 points on each graph and each point is (accuracy SVM/GP, accuracy Mixture GP) and corresponds to one test run. Circle width is proportional to the number of points having that coordinate. Points above the diagonal indicate where Mixture GP was more accurate. While Mixture GP is better (a) or comparable (c) on average when all the modes are intact, it is particularly better when there are noisy and missing channels (b) and (d).

Once we have the posterior over the switches, $Q(\lambda_i)\ \forall i \in [1..n]$, we first infer the switches for the test data $\bar{\mathbf{x}}^*$ using a meta-level GP based classification system. For this, we do a multi-label P-way classification using the GP algorithm described in 4.2 with $\hat{\mathbf{\Lambda}} = \arg\max_{\mathbf{\Lambda}} Q(\mathbf{\Lambda})$ as labels. Specifically, for an unlabeled point $\bar{\mathbf{x}}^*$, P different classifications are done where each classification provides us with q_r^*, where $r \in \{1, .., P\}$, and equals the probability that channel r was chosen to classify $\bar{\mathbf{x}}^*$. The posterior $Q(\lambda^* = r)$ is then set to $\frac{q_r^*}{\sum_{p=1}^{P} q_p^*}$. In our experiments, to perform this multi-label P-way classification, we clubbed all the channels together using -1 as observations for the modalities that were missing. Note, that we are not limited to using all the channels clubbed together; but, various combinations of the modalities can be used including other indicator and contextual variables. Once we have the posterior over the switch for the test data, $Q(\lambda^*)$, we can infer the class probability of an unlabeled data $\bar{\mathbf{x}}^*$ using:

$$p(t^*|\bar{\mathbf{X}}, \mathbf{t}) = \int_{\bar{\mathbf{Y}}, \lambda^*} p(t^*|\bar{\mathbf{Y}}, \lambda^*) Q(\lambda^*) Q(\bar{\mathbf{Y}}) \tag{4}$$

The main feature of the algorithm is that the classification using EP is required only once and the bound in equation 3 can be optimized very quickly using the Gaussian approximations provided by EP. For details please refer to [5].

5. EMPIRICAL EVALUATION

We performed experiments on the collected database using the framework to classify the state of interest (65 samples) vs. uninterest (71 samples). The experimental methodology was to use 50% of the data for training and use the rest for testing. Besides the comparison with the individual modalities, we also compare the mixture of GP with the HMM based expert-critic scheme [8] and a naive feature level fusion. In the naive feature level fusion, the observations from all the channels are stacked to form a big vector and these vectors of fused observations are then used to train and test the classifiers. However, in our case this is not trivial as we have data with missing channels. We test a naive feature level fusion where we use -1 as a value of all those observations that are missing, thus, fusing all the channels into one single vector.

All the experiments were done using GP classification as the base classifiers and for completeness we also perform comparisons with SVMs. Both, GP classification and SVMs used RBF kernels and the hyperparameters for GP classification (σ, β) were selected by evidence maximization. For SVMs we used the leave-one-out validation procedure to select the penalty parameter C and the kernel width σ. We randomly selected 50% of the points and computed the hyper-parameters for both GPs and SVMs for all individual modalities and the naive feature level fusions. This process was repeated five times and the mean values of the hyperparameters were used in our experiments. Table 2 shows the

Table 2: Recognition rates (standard deviation in parenthesis) averaged over 100 runs.

	GP	SVM
Upper Face	66.81%(6.33)	69.84%(6.74)
Lower Face	53.11%(9.49)	57.06%(9.16)
Posture	81.97%(3.67)	82.52%(4.21)
Game	57.22%(4.57)	58.85%(5.99)
Mix of GP	**86.55%(4.24)**	-

results for individual modalities using GP classification and SVMs. These numbers were generated by averaging over 100 runs and we report the mean and the standard deviation. We can see that the posture channel can classify the modalities best, followed by features from the upper face, the game and the lower face. Although, the performance obtained using GP classification is similar to SVMs and slightly worse for upper and the lower face, we find that extension of GP to a mixture boosts the performance and leads to significant gains over SVMs. The Mix of GP also outperformed fixed rule based decision fusion of the individual SVM and GP classifiers. The readers are requested to look at [5] for similar results.

Next, we compare the Mix of GP to feature level fusion while restricting the training and testing database to the points where all the channels are available. The restriction results in a significant decrease in the available data (only

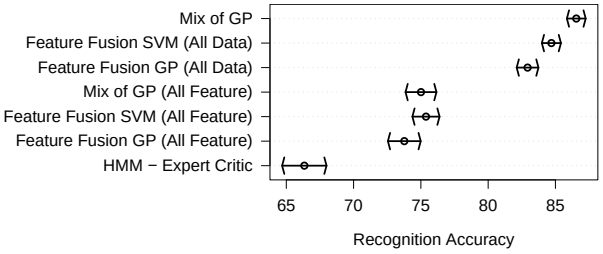


Figure 6: Performance comparison of the proposed Mixture of GP approach on the affect dataset with naive feature level fusions and the HMM based expert-critic framework. Each point was generated by averaging over 100 runs. Non overlapping of error bars, the standard errors scaled by 1.64, indicates 95% significance of the performance difference.

50 total datapoints) and the mix of GP obtained an average accuracy of 75.01% ± 1.15 over 100 runs, whereas the accuracy using feature level fusion was 73.77%±1.22. When we perform naive feature level fusion using all the data, without any restriction and using -1 for the channels that were missing, we get a significant gain in average accuracy to 82.93%±0.81. However, the Mix of GP outperforms all of these significantly with an average accuracy of 86.55%±0.55 (see figure 6 for significance). Figure 5(a) and (b) show that the Mixture of GP not only beats feature level fusion on the subset of data where all channels are present, but also significantly beats it when incomplete data can be used. Similar results are obtained comparing to SVMs and are graphically shown in figure 5 (c) and (d). Figure 6 graphically demonstrates the performance gain obtained by the Mix of GP approach over the feature level fusions and the HMM based expert critic framework as implemented in [8] with FACS features. The points in figure 6 were generated by averaging over 100 runs. The non-overlapping error bars, standard errors scaled by 1.64, signify 95% confidence in performance difference. The Mix of GP uses all the data and can handle the missing information better than naive fusion methods; thus, it provides a significant performance boost.

In our earlier work [8, 5], we had used manually encoded FACS based Action Units (AU) as features extracted from the face. The accuracy of 66.81 ± 6.33 obtained while performing GP classification with automatically extracted upper facial features was significantly better than the accuracy of 54.19 ± 3.79 obtained with GP classification that used the manually coded upper AUs. The difference is due to two factors. First, the set of automatically extracted upper facial features is richer than the AUs. Second, the AUs were manually encoded by just one FACS expert, thus, resulting in features prone to noise.

6. CONCLUSION

In this paper, we proposed a Mixture of Gaussian Processes approach to classifying interest in a learning scenario using multiple modalities under the challenging conditions of missing channels and uncertain labels. Using information from upper and lower face, postures and task information, the proposed multisensor classification scheme outperformed several sensor fusion schemes and obtained a recognition rate of 86%. Future work includes incorporation of active learning and application of this framework to other challenging problems with limited labeled data.

Acknowledgments

Thanks to Selene Mota for help with the data collection. This research was supported by NSF ITR grant 0325428.

7. REFERENCES

[1] T. W. Chan and A. Baskin. *Intelligent Tutoring Systems: At the Crossroads of Artificial Intelligence and Education*, chapter 1: Learning companion systems. 1990.

[2] C. Conati. Probabilistic assessment of user's emotions in educational games. *Applied Artificial Intelligence, special issue on Merging Cognition and Affect in HCI*, 16, 2002.

[3] T. S. Huang, L. S. Chen, and H. Tao. Bimodal emotion recognition by man and machine. In *ATR Workshop on Virtual Communication Environments*, 1998.

[4] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991.

[5] A. Kapoor, H. Ahn, and R. W. Picard. Mixture of gaussian processes to combine multiple modalities. In *Workshop on MCS*, 2005.

[6] A. Kapoor, S. Mota, and R. W. Picard. Towards a learning companion that recognizes affect. In *AAAI Fall Symposium*, Nov 2001.

[7] A. Kapoor and R. W. Picard. Real-time, fully automatic upper facial feature tracking. In *Automatic Face and Gesture Recognition*, May 2002.

[8] A. Kapoor, R. W. Picard, and Y. Ivanov. Probabilistic combination of multiple modalities to detect interest. In *ICPR*, August 2004.

[9] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *PAMI*, 20(3):226–239, 1998.

[10] D. J. Miller and L. Yan. Critic-driven ensemble classification. *Signal Processing*, 47(10), 1999.

[11] T. P. Minka. Expectation propagation for approximate bayesian inference. In *UAI*, 2001.

[12] S. Mota and R. W. Picard. Automated posture analysis for detecting learner's interest level. In *CVPR Workshop on HCI*, June 2003.

[13] N. Oliver, A. Garg, and E. Horvitz. Layered representations for learning and inferring office activity from multiple sensory channels. In *ICMI*, 2002.

[14] M. Pantic and L. J. M. Rothkrantz. Towards an affect-sensitive multimodal human-computer interaction. *Proceedings of IEEE*, 91(9), 2003.

[15] R. W. Picard, E. Vyzas, and J. Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *PAMI*, 2001.

[16] K. Toyama and E. Horvitz. Bayesian modality fusion: Probabilistic integration of multiple vision algorithms for head tracking. In *ACCV*, 2000.