A Robust Joint Face Model for Human Emotion Recognition

Ayesha Hakim
School of Engineering and Advanced Technology (SEAT)
Massey University
New Zealand
a.hakim@massey.ac.nz

Stephen Marsland
School of Engineering and Advanced Technology (SEAT)
Massey University
New Zealand
s.r.marsland@massey.ac.nz

Hans W. Guesgen
School of Engineering and Advanced Technology (SEAT)
Massey University
New Zealand
h.w.guesgen@massey.ac.nz

ABSTRACT

Having computers recognise emotions has benefits for human-computer interaction, psychology, and behavioural analysis. Unfortunately it is a very difficult problem, partly because humans can, to some extent, control the appearance of an emotion on their faces. In this paper we show that building statistical shape models of different parts of the face and combining them can give more successful results than using only a model of the whole face.

Categories and Subject Descriptors

I.4.9 [**Computing Methodologies**]: Image Processing and Computer Vision—*applications*

General Terms

Algorithms

Keywords

Emotion recognition, Principal Component Analysis, Support Vector Machine, Mahalanobis Distance

1. INTRODUCTION

The desire to make human-computer interactions as natural as possible is nearly as old as computers, but has met with little success so far. One area that is receiving significant current interest is enabling machines to recognise human emotions based on analysis of images from cameras. In this paper we introduce a method of classifying emotions based on statistical models of both the full face and, separately, the upper and lower halves of the face. We demonstrate that this gives more reliable classification of the basic emotions than standard methods described in the literature.

The term 'emotion' is a rather vague one, making emotion recognition and understanding a difficult problem. Psychologists, physiologists and philosophers have debated definitions of the term for many years, but we still do not have a standard definition [13]. For computers, a significant part of the problem is that emotions are internal, and hence impossible to see directly. However, there are generally accepted physical correlates, principally facial expression and tone of voice, although brain activations can also be correlated with emotion [13]. In common with many other researchers, we will assume that when we feel an emotion, it appears on our face [4]. This enables machine vision tools to be used to identify emotions, even though the emotions themselves are internal, private feelings.

There has been a lot of research into emotion recognition, analysis, and synthesis over the past three decades. Where these approaches are based on machine vision, they typically work on static images of the face, as we do in this paper. The extension to video analysis, which has the added benefit of enabling time series analysis of emotions, is discussed at the end of the paper. Approaches that have been explored in the literature include methods for relating face images to physical structure of the facial skin and musculature [7], measurements of the shapes of facial features and their spatial arrangements [9], gray level pattern analysis using local spatial filters [15], and holistic spatial pattern analysis using techniques based on principal component analysis [9, 15].

Another common method for measuring facial expressions in behavioural science is the Facial Action Coding System (FACS) [6]. FACS is a scoring system defined for expert human observers, not computers. It aims to provide objective measures of facial activity to assist behavioural science analysis of the face. Ekman and Friesen defined 46 Action Units (AUs) that described the smallest visually distinguishable facial movements. A lot of researchers have tried to automate FACS by recognizing facial action units in images and/or videos, e.g. [10, 11, 16]. Although FACS is a promising approach, in reality it is not always possible to locate the action units in each image/frame due to changes in lighting conditions, occlusions caused by the facial hairs and glasses, and poor image quality.

2. DATA PREPROCESSING

While analysis of images of the human face for emotions is the ideal approach, we have chosen to start with a simpler problem. We have selected a dataset (the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset from the Speech Analysis and Interpretation Laboratory (SAIL) at the University of South California (USC) [2]) where actors were recorded during both scripted and improvised sessions with markers on their face and hands. High speed cam-

IVCNZ '12, November 26 - 28 2012, Dunedin, New Zealand

eras (120 frames per second) were used to record the actors, and the 3D location of the points marked on the face were recorded at this speed. The released version of the dataset contains sets of points for fourteen improvised and fourteen scripted conversations lasting approximately 2 minutes each between two actors, one male and one female. For each conversation, only one of the two actors was recorded.

Within each conversation, each utterance of the two speakers (sentence or similar period during which one actor talks) was annotated by three independent human evaluators into categorical labels (neutral, happy, angry, sad, surprise, disgust, fear, frustration, and excitement) as well as psychological data about emotion intensity (valence, activation, and dominance). For further details on this, see [2].

We used the locations of the marker points in 3D as the basis for our analysis, and randomly chose 4,000 frames of each of six emotions (neutral, happy, excitement, angry, frustration, and sad) to form a training set of 24,000 frames, creating a separate set for each of the two actors. The labelling of each frame chosen was based on the three expert annotations. It was rare to find all three experts assigning the same label, so we took frames where at least two of them agreed. We used six emotions rather than the full nine as for the missing emotions (disgust, surprise, and fear) there was insufficient data, sometimes as little as 2,000 frames in total. We also selected 16,000 frames for each subject to form a test set. The order of the frames was randomized, and each frame was analyzed independently.

Each frame of the dataset contains the motion capture information of 55 markers in 3 dimensions, so the training data is of size $24,000 \times 165$ dimensions. We reduced the dimensionality of the data for each frame in three ways:

1. Markers not on the face (such as the hands) were excluded.
2. Markers that did not move significantly (such as eyelids and nose) were removed.
3. Sets of markers that moved together (such as points on the chin and forehead) were replaced by a single point at the centre of the set.

As a result of these simplifications each emotion frame was represented by 28 markers in 3D, which is an 84D vector.

3. SHAPE MODELS

A standard technique to analyze sets of datapoints in high dimensional spaces is Principal Component Analysis (PCA). This has been used many times for face recognition and emotion analysis, such as eigenfaces [19] and Statistical Shape Models [3]. PCA identifies a linear transformation (translation and scaling) of the dataset so that the resultant basis has maximum variance of the dataset along the first basis vector, and successively less variance amongst the following (mutually orthogonal) basis vectors. It is a commonly used method for dimensionality reduction [12].

We applied PCA to each 84D datapoint. The first five principal components (PCs) covered 93% of the total variation of the training data. We then examined the effect of each PC independently, by applied changes of ± 3 standard deviations from the mean face along each PC. We noticed that the first PC, which covers 48.4% of the total variation, was contributing to the upward and downward movement of the mouth points (i.e. lips). We hypothesised that this was due to talking, and therefore selected a set of silent frames from the data and applied PCA to these points. Since the first PC was not present in the silent frames, we decided that it was primarily correlated with talking and therefore not directly connected with emotion recognition. We therefore discarded it. In fact, for emotion recognition, talking is often found to be one of the biggest confusion factors and sources of error [18].

Based on this we chose to use four PCs (2-5) for our analysis. We took the training data and transformed it into the 4D space of these four principal components. Each datapoint was then labelled with the majority vote of the three experts, so that we had 24,000 points with one of six labels in a 4D space.

3.1 Classification

For classification we replaced each cluster (i.e., set of points labelled as one emotion) with the mean of that set, and also computed the covariance matrix (spread) of the data. We therefore ended up with six datapoints representing the mean of each set and an associated covariance matrix. For an (unknown) testing frame we transformed it into the 4D space and then computed the Mahalanobis distance between it and each of the six cluster centres. We then labelled that test point with the label of the cluster that it is closest to. The Mahalanobis distance uses the spread to compute a distance that includes the amount of uncertainty in the data. It is formulated as:

$$D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

where x is the (4D) column vector of the testing frame, μ is a column vector of the mean and Σ is the 4×4 covariance matrix for an emotion.

3.2 Using the Upper and Lower Face Separately

According to [1], some expressions are better recognized from muscle activity in the upper half of the face, while others use muscles primarily from the lower half of the face. Moreover, [5] suggested that the upper face (including the eyebrows and forehead) is difficult to control voluntarily as compared to the lower face including the mouth, cheeks and chin. If this is true, then the model based on just the upper face marker points would be more reliable and would lead to true classification of emotions. On the basis of these observations, we also created separate shape models of the upper and lower halves of the face. With the IEMOCAP dataset the upper-half contains 17 marker points covering the forehead, eyebrows, and eyes (Figure 3); while the lower-half consists of 11 marker points covering the cheeks, lips, and chin (Figure 4). For the upper face, the first 5 PCs were selected, covering 93.4% of the total variance and for the lower face the first 4 PCs were selected covering almost 95% of the total variance of the data.

Based on this we have three separate shape models of the dataset: the full face, and the upper and lower faces separately. We chose to combine the three models by calculating the distance between the testing frame and the transformed training set of each emotion by the model using the Mahalanobis Distance (Figure 1). In this way, we get three sets of distances; one for each model. Each set contains distance of testing frame from the six emotions. Finally, the

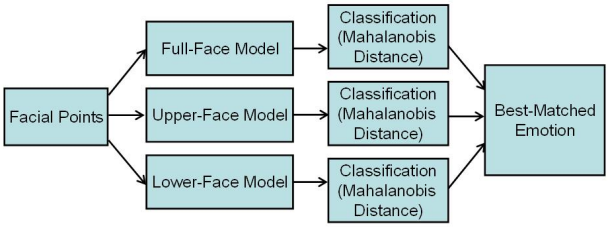


Figure 1: The Proposed Joint Face Model

testing frame is classified as the emotion which is closest to that frame in terms of Mahalanobis Distance. Since all four models are 4D this should be a reasonable estimate.

4. THE EFFECTS OF EACH PRINCIPAL COMPONENT

Having already identified that PC1 was involved in talking, we chose to investigate the effects of the other four PCs of the full model as well. PC2 covers 28% of the total variation of data. It controls the sideways movement of mouth points (lips). The outward movement of the lips contribute to the positive expressions (e.g. smile) while the inward movement gives the negative expressions (e.g. sad). The third PC covers 7.8% of total data variation and contributes to the upward and downward, while PC4 covers 6% of total variation and contributed to the sideways movement (i.e. towards and away from the nose) of eyebrows and forehead marker points. The fifth PC covers 3% of the variation of data and appears as a rather strange circular motion of the lips. Experimentally we observed that this PC contributed mainly to the laughing expression.

The following table describes the effects, which are the same for both the female and male shape models, while Figure 2 shows the effects of varying PC2 for the female model. In the table, '+' means moving computing $\mu + \lambda\sigma_i$ for positive weights λ (where μ is the mean datapoint and σ_i is the ith principal component, while '-' means using negative weights λ).

PC2	+	Outward movement of lips
PC2	−	Inward movement of lips
PC3	+	Upward movement of eyebrows & forehead
PC3	−	Downward movement of eyebrows & forehead
PC4	+	Inward movement of eyebrows & forehead
PC4	−	Outward movement of eyebrows & forehead
PC5	+	Right to left circular movement of lips
PC5	−	Left to right circular movement of lips

4.1 Effect of each Full, Upper, and Lower Face PC

For the *upper-face*, the first PC covers almost 46% variation. Since this model cannot see the effects of talking, this PC identifies upward motion of the eyebrows and forehead, while the second PC (figure 3) also deals with similar movement and covers 34.6% of the variation. The third PC, which covers 5.5% of the variation highlights the inner brows moving upward, the fourth PC (4.3%) is the outer brow moving upward and the fifth PC identifies the right brow moving downward and the left brow moving in the upward direction.

In the case of *lower-face* model, the first PC covers 56.5% variation of the data identifies the upward and downward movement of mouth and the second PC (figure 4) covers 27.5% variation deals with the sideways movement of mouth. The first and second PCs of the lower face are same as that of the full face PCs. The third PC covers 7.6% of the variation highlights the upward movement of chin and cheeks, while the fourth PC covers 3.2% of the variation identifies the circular movement of lips similar to the fifth PC of the full face.

It should be noted that all the figures are actually in 3D, but to get a clear view the viewpoint has been rotated by setting the azimuth and elevation equal to zero.

5. PERFORMANCE COMPARISON

Based on the shape models we had four methods of classifying emotions: using the joint model, or any one of the three models separately. As a comparison, we implemented two methods that have been used in the past: a set of rule-based classifiers [8] and a set of Support Vector Machines (SVM) [14]. These methods are described in this section.

5.1 A Rule-Based Emotion Classifier

In the previous section we looked at the effect of each principal component separately. This enabled us to create a simple rule-based classifier, where the rules we applied were:

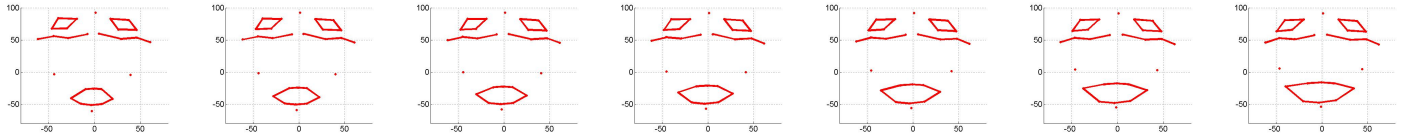
Happy	*PC2* is positive and larger than *PC3*
Excited	*PC2* is positive but *PC3* is larger
Laughing	*PC2* is positive but *PC5* is larger
Angry	*PC2* is negative, *PC3* is positive, and *PC4* is negative
Frustrated	*PC2* is negative, *PC3* is negative, and *PC4* is negative
Sad	*PC2* is negative, *PC3* is negative, and *PC4* is positive

These rules also highlight something that is known from psychology, and that we observed in our data, namely that angry and frustrated look similar on the face, as do happy and excited. These clusters of points overlap significantly. According to these rules, angry and frustrated shares the negative effect of PC2 and PC4 and happy and excited both have the positive effect of PC2. Happy and angry are well separated, while sad and angry also overlap some points since both have the negative effect of PC2.

5.2 Support Vector Machines

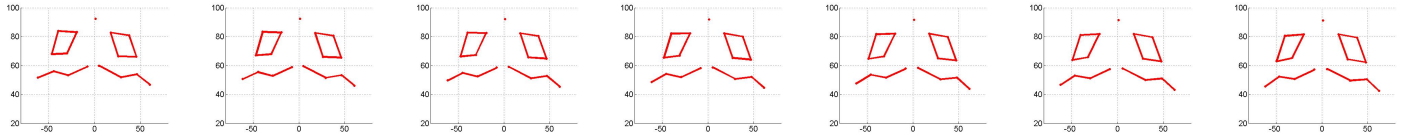
The Support Vector Machine (SVM) is a popular machine learning technique for emotion recognition [14]. We used an SVM with a quadratic kernel using Sequential Minimal Optimization (SMO) method [17], implemented using Matlab's svmclassify from the Bioinformatics toolbox. The SVM is a binary classifier, so we convert it to a multi-class classifier by using one-versus-all training, where a separate SVM is trained to recognise each class, compared to all of the others. As a result, we obtained six SVMs, one for each emotion.

Each SVM is trained on the original training data without transformation with the labels where at least two human observers agreed. Classification was then performed on the basis of class membership (± 1), where $+1$ means successfully matched to the training class.



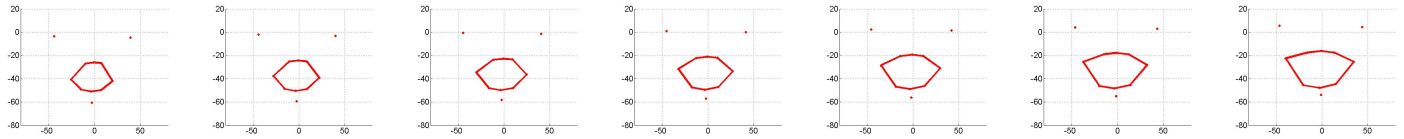
(a) $\mu - 3\sigma$ (b) $\mu - 2\sigma$ (c) $\mu - 1\sigma$ (d) μ (e) $\mu + 1\sigma$ (f) $\mu + 2\sigma$ (g) $\mu + 3\sigma$

Figure 2: The effects of PC2 for the female full face model.



(a) $\mu - 3\sigma$ (b) $\mu - 2\sigma$ (c) $\mu - 1\sigma$ (d) μ (e) $\mu + 1\sigma$ (f) $\mu + 2\sigma$ (g) $\mu + 3\sigma$

Figure 3: The effects of PC2 for the female upper face model.



(a) $\mu - 3\sigma$ (b) $\mu - 2\sigma$ (c) $\mu - 1\sigma$ (d) μ (e) $\mu + 1\sigma$ (f) $\mu + 2\sigma$ (g) $\mu + 3\sigma$

Figure 4: The effects of PC2 for the female lower face model.

6. RESULTS

In this section we report two experiments: first comparing the four shape models, the rule-based classifier, and the SVM-based classifier based on training with 24,000 datapoints and either six or four classes, and then examining the robustness of the more successful methods to mislabelling of data.

The six classes that we used in our dataset were neutral, angry, frustrated, happy, excited and sad. However, we noticed that angry and frustrated had a significant overlap, as did happy and excited. This has also been reported in the literature [2]. We therefore also used four classes: neutral, angry/frustrated, happy/excited, and sad, which is a significantly easier problem.

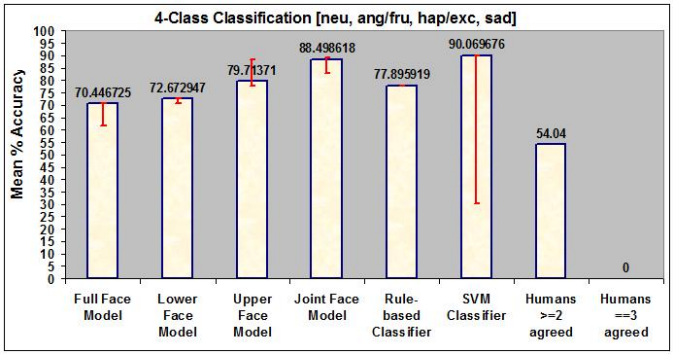


Figure 5: The comparison of all model's performance by using 4-classes [neu, ang/fru, hap/exc, sad] on female Dataset

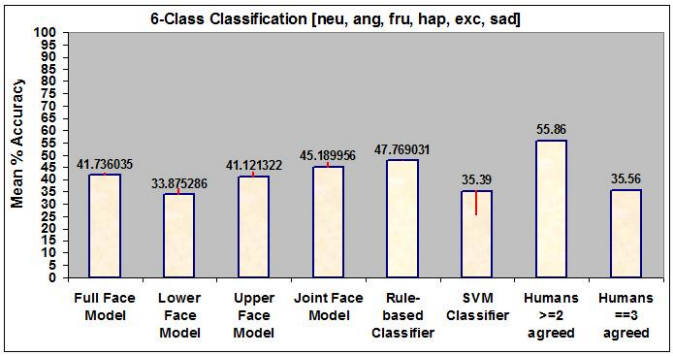


Figure 6: The comparison of all model's performance by using 6-classes [neu, ang, fru, hap, exc, sad] on male dataset

Figure 5 shows the results on all 16,000 test images for the six automatic methods of emotion recognition as well as for agreement between the three human observers for the female dataset. While in the training set only images where at least two of the experts agreed were used, the test set contains images where all three provided different assessments, often because one or more of them gave two labellings (such as angry/frustrated). The label assigned was the maximum of the more than three labels attached to the frame by the three experts. It can be seen that the *joint-model* returns almost 89% accuracy, which is consistently better than the *full* (70.45%), *upper* (79.71%), and *lower* (72.67%) face models, as well as *rule-based* PCA classifier (77.89%). On the original female dataset (with no mislabelled frames) SVM seems to outperform other models, but it does not show consis-

	Female 4-class mean(std)	Female 6-class mean(std)	Male 4-class mean(std)	Male 6-class mean(std)
Joint Face Model	88.50(0.59)	32.10(1.33)	65.51(1.55)	45.19(1.76)
Full Face Model	70.45(8.655)	22.64(0.46)	59.34(6.4)	41.74(0.88)
SVM Classifier	90.07(59.72)	38.55(17.30)	35.46(8.06)	35.39(9.77)

Table 1: Mean accuracy and standard deviation of the joint face model, full face model and SVM Classifier on both 4 and 6 emotion classes

tent performance and remains unpredictable specially in the tough data with high level of mislabelling. In order to compute the difference in the *means* of joint model and SVM, we perform the *t-Test: Paired Two Sample for Means* [21] on the results of these two models. The test suggests that there is no significant difference between these two results ($p > 0.05$). Figure 6 shows the results of classification to six emotion classes as well as for agreement between the three human observers for the male dataset. The classification to the six emotion classes is a difficult problem due to the significant overlap of angry and frustrated as well as the happy and excited data clusters. Table 1 lists the mean accuracy and standard deviation of the joint face model, full face model and SVM Classifier on both 4 and 6 emotion classes. It shows that the joint face model significantly improves the accuracy as compared to other methods.

The proposed joint model performs better than the human observers despite the fact that the human observers had a whole lot of other information like voice, head, and hand motion along with the facial expressions, while the system has to classify on the basis of just the motion of facial markers.

6.1 Robustness to Mislabelled Data

There are two problems with the training data. First, as has just been observed, the human experts were inconsistent and we were worried that the labels of the training data were compromised by errors. However, part of the problem may well be that they only labelled utterances, and each utterance had many frames within it. The emotion label was attached to all frames in the utterance, but there is no guarantee that they all show one consistent emotion. In order to test the robustness of our system to data mislabelling, we manually mislabelled some proportion of the training data and ran the experiment again. The mislabelling was performed by choosing frames at random and then changing their label. This was done keeping in mind the possible errors that were likely to occur, so angry could become frustrated, frustrated could become angry or sad, neutral could become sad, and so on. We created datasets where 0.25%, 0.5%, 1.25%, 2.5%, 5%, 17.5%, 25%, 50% and 100% of the data were mislabelled.

We tested and compared all above mentioned models including *full*, *upper*, and *lower* face models; proposed *joint-model*; *rule-based* PCA classifier and *SVM* classifier on gradually increasing mislabelled training data using both 6 and 4 emotion classes. Figure 7 presents the performance of proposed joint model and SVM classifier on the female dataset. The graphs show that the system is resistant until 25% of the training data (i.e. 1000 out of 4000 frames) is mislabelled for the joint model, however the performance of SVM is not consistent.

7. CONCLUSIONS AND FUTURE WORK

Giving a discrete name to emotion is not always helpful, since in daily life we come across a lot of complex emotional states that can not be given a discrete name. In other words, sometimes the difference between two or more emotions is so subtle that it becomes difficult to finely differentiate between them.

This research is an intermediate step towards our goal of representing the emotions in a lower-dimensional emotion space. In future work, instead of classifying the emotions based on the best-matched criteria, we would be adding the effect of all basic emotions for a better representation of complex emotions. In psychology, activation-evaluation space is the most widely used method of representing emotions [20]. The distance and location of emotional points in this space may give some idea about the valence and activation of basic as well as complex emotional states.

Moreover, although the first PC which contributes to the talking movement has been discarded in emotion recognition, but it may be useful in applications like speaker detection (whether a person is speaking or not). In future, we would be combining the applications of face recognition along with the proposed speaker detection using mouth model for speaker recognition (who is speaking).

8. REFERENCES

[1] J. N. Bassili. Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology*, 37:2049–2058, 1979.

[2] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Journal of Language Resources And Evaluation*, 4(42), 2008.

[3] A. J. Calder, A. Burton, P. Miller, A. W. Young, and S. Akamatsu. A principal component analysis of facial expressions. *Vision Research*, 41(9):1179 – 1208, 2001.

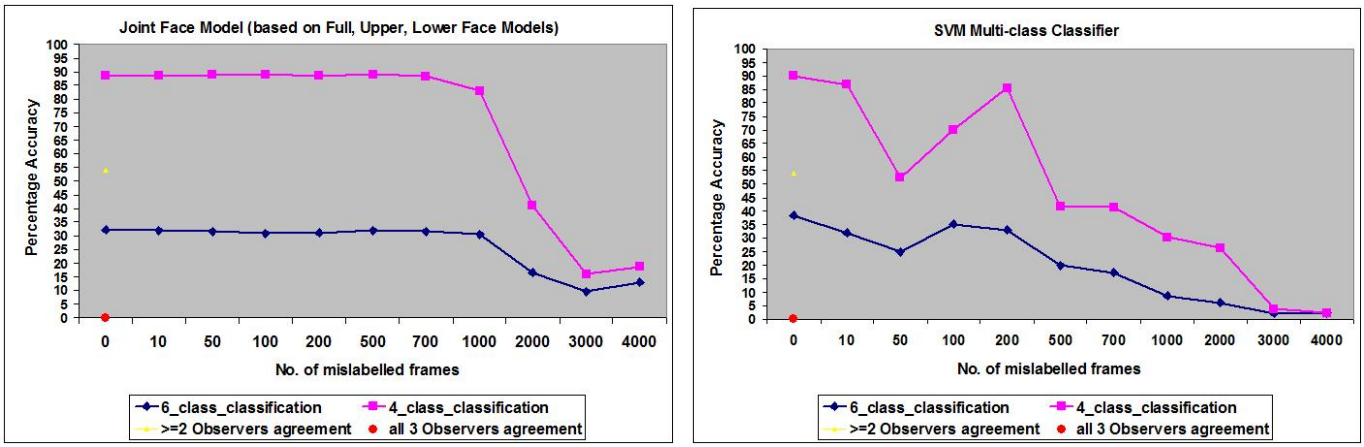
[4] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 2(17):124–129, 1971.

[5] P. Ekman and W. V. Friesen. *Unmasking the face*. Englewood Cliffs, N.J.: Prentice-Hall, 1975.

[6] P. Ekman and W. V. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Action*. Palo Alto: CA: Consulting Psychologists Press, 1978.

[7] I. A. Essa and A. P. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:757–763, 1997.

[8] S. V. Ioannou, A. T. Raouzaiou, V. A. Tzouvaras,



(a) The Joint Face Model

(b) SVM Classifier

Figure 7: The robustness of Joint Face Model and SVM to mislabelled training data

T. P. Mailis, K. C. Karpouzis, and S. D. Kollias. Emotion recognition through facial expression analysis based on a neurofuzzy network. *Journal of Neural Networks*, 18:423 – 435, 2005.

[9] A. Lanitis, C. J. Taylor, and T. F. Cootes. Automatic interpretation and coding of face images using flexible models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:743–756, 1997.

[10] G. C. Littlewort, M. S. Bartlett, and K. Lee. Automatic coding of facial expressions displayed during posed and genuine pain. *Journal of Image and Vision Computing*, 27(12):1741–1844, 2009.

[11] M. H. Mahoor, S. Cadavid, D. S. Messinger, and J. F. Cohn. A framework for automated measurement of the intensity of non-posed facial action units. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 1 and 2:833–839, 2009.

[12] S. Marsland. *Machine Learning: An Algorithmic Perspective*. Chapman and Hall CRC, 2009.

[13] G. Matthews, M. Zeidner, and R. D. Roberts. *Emotional intelligence: Science and myth*. Cambridge, MA: MIT Press, 2002.

[14] P. Michel and R. Kaliouby. Real time facial expression recognition in video using support vector machines. In *IEEE International Conference on Multimodal Interfaces (ICMI)*, pages 258 – 264, 2003.

[15] C. Padgett and G. Cottrell. Representing face images for emotion classification. *Advances in Neural Information Processing Systems*, 9:894–900, 1997.

[16] M. Pantic and L. J. M. Rothkrantz. An expert system for recognition of facial actions and their intensity. In *Seventeenth National Conference on Artificial Intelligence (Aaai-2001) / Twelfth Innovative Applications of Artificial Intelligence Conference*, pages 1026–1033, 2000.

[17] J. C. Platt. *Fast training of support vector machines using sequential minimal optimization*. MIT press, 1999.

[18] N. Sebe, I. Cohen, T. Gevers, and T. Huang. Emotion recognition based on joint visual and audio cues. In *International Conference on Pattern Recognition*, page 1136Ű1139, 2006.

[19] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1), 1991.

[20] C. Whissell. *The dictionary of affect in language*, volume 4. The Measurement of Emotions: Academic press, New York, 1989.

[21] J. Zar. *Biostatistical analysis*. Prentice Hall International, 3 edition, 1996.