Facial Emotions and Emotion Intensity Levels Classification and Classification Evaluation

Ing. Marián Beszédeš
Faculty of Electrical Engineering
and Information Technology
Slovak University of Technology
Bratislava, 81219, SK
beszedes@ktl.elf.stuba.sk

Dr. Phil Culverhouse
School of Computing,
Communications and Electronics
University of Plymouth
Plymouth, PL48AA, UK
P.Culverhouse@plymouth.ac.uk

Abstract

In this paper we analyze the problem of human facial emotion and emotion intensity levels recognition and resulting classification accuracy evaluation. Final testing set classification accuracy value is usually taken as a quantifier of method quality. However, this value is often strongly affected by the testing set parameters such as number, age and gender of subjects or intensity of their emotions etc. In this work we propose a different classifier evaluation methodology that uses the human visual system as a reference point. We employed active appearance models and support vector machines for facial emotion classification. Our SVM classifier gave slightly more consistent labels to emotion categories for images than human subjects, while humans were more consistent at identifying emotion intensity level than SVM.

1 Introduction

Automatic facial expression analysis has become an interesting research area since the early 90's because of many potential applications in areas like human-computer interfaces, face appearance synthesis, image retrieval and human emotion analysis.

Ekman and Friesen [6] has postulated six primary emotions which seem to be universal across human ethnicities and cultures. This primary emotions are commonly referred as *basic emotions* and comprise happiness, anger, surprise, disgust, fear and sadness. Despite many questions that arise around this study [12] (are the basic emotions indeed universal? Does a facial expression have a strong relation with an actual emotion state ?) it is still widely used in computer vision community.

While the human capability for face detection is very robust, the same has not been proven for facial expressions interpretation. According to Bassili [1], a person trained to classify faces can recognize six basic emotions with an average accuracy 87%. This accuracy ratio can change due to several reasons like the familiarity with the face, general experience with different types of expression, the facial emotion intensity level or even subjects or recognizing person race. For example the smiling face with low intensity can be easily misinterpreted as a neutral facial expression. Most current works related to automatic facial emotion recognition deal just with facial emotions having the highest

BMVC 2007 doi:10.5244/C.21.52

intensity. The classification of emotion intensity levels is in most of the cases omitted too. Surveys with more details related to analysis of facial expressions and emotions can be found in [7] or [11].

Facial emotion recognition in still images is connected with three basic tasks: 1) face detection, 2) facial expression data extraction and finally, 3) facial classification. There have been many techniques for face detection invented in last 15 years. For example the Viola-Jones face detector based on Haar feature classification and AdaBoost algorithm [13] combined with accurate skin color detector [2] can solve this complex task in real-time with very good classification accuracy. Fiducial grid and Gabor wavelets [9] or PCA analysis of random patches [10] are just examples of many other methods that were employed for the 2nd task of facial expression data extraction. Active Appearance Models (AAM) method is another well known method for accurate facial feature detection and facial emotion features extraction is approach [5],[8]. It can be relatively easily implemented, facial features are detected very fast and precisely and it yields good results on difficult and noisy data. Finally, the 3rd task of extracted facial data classification can be resolved using many types of classifiers. PCA, LDA and Mahalonobis distance [5],[9], neural network based classifiers [10] or SVM classifiers [8] methods were utilized for this aim.

The above methods are usually assessed on testing sets after all training procedures. The value that stands for portion of correctly classified samples from testing set is taken then as *final classification accuracy* κ and quantifier that tells about qualities of used method. However, κ is often strongly affected by the testing set parameters. Number, age, race, and gender of subjects; number, type and intensity of emotions; quality of images, lighting conditions; all these factors can impact the classification results. For example, in [8] Liebelt recognized 7 facial emotions with $\kappa = 71.3\%$ and no details about testing set parameters were given. Lyons used set of 193 images showing 9 Japanese females and 7 facial emotions for training in [9]. $\kappa = 75\%$ was achieved for novel subjects but no details about this kind of testing set were given again. 84 Ekmans' facial emotion photos with 12 subjects and 7 emotions were used in the work of Padget [10]. They trained their classifier on images of 11 subjects and tested on the images of the 12th subject. By changing the training and testing set they get average $\kappa = 86\%$.

There is no commonly used database of facial images with varying emotions as can be seen from previously mentioned works. One of the reasons of this lack is that researchers usually need images of different quality, lighting conditions, contrast, bit depth of colors etc. If we want to evaluate the classification results without considerations of testing set parameters, a reference classifier is needed: such as the human visual system.

In this paper we describe a classification system based on Active Appearance Models (AAM) method which is used for facial emotion features extraction and Support Vector Machines (SVM) method, which is used for classifiers training. An outline of the theoretical basis of the AAM model and the way we use it for proper feature extraction is presented. Then, important SVM classifier attributes are discussed as well as procedures suitable for its training. Our machine methods are referenced to a study that used human participants. The experimental methodology is presented and results of statistical analysis discussed in the context of the machine performance data.

2 Active Appearance Models

AAM learn characteristics of objects during a training phase by building a compact statistical model representing shape and texture variation of objects. As it is described in [4], the concept AAM is based on the idea of combining both shape and texture information of the objects to be modeled. The appearance model is built based on a set of N labeled images, where n key landmark points are marked on each example object and form so called shape vectors. The shape vectors $\mathbf{x}^i = (x_1^i, \ldots, x_n^i; y_1^i, \ldots, y_n^i)^T, i = 1 \ldots N$ are aligned using Procrustes analysis and the labeled images are warped to the mean shape $\bar{\mathbf{x}}$ and normalized, yielding the texture vectors $\mathbf{g}^i$. By applying principal component analysis (PCA) to the normalized data, linear models are obtained for both shape, $\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}_s\mathbf{b}_s$, and texture, $\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P}_g\mathbf{b}_g$, where $\bar{\mathbf{x}}$, $\bar{\mathbf{g}}$ are the mean vectors, $\mathbf{P}_s$, $\mathbf{P}_g$ are sets of orthogonal modes of variation (the eigenvectors resulting from PCA), and $\mathbf{b}_s$, $\mathbf{b}_g$ are sets of model parameters.

A given object can thus be described by $\mathbf{b}_s$ and $\mathbf{b}_g$. As $\mathbf{P}_s$, $\mathbf{P}_g$ may still be correlated, one more PCA is applied to them. This yields the final combined linear model $\mathbf{b} = \mathbf{P}_c\mathbf{c}$, where $\mathbf{P}_c = (\mathbf{P}_{cs}\mathbf{P}_{cg})^T$.

The goal of AAM search is to find the model parameters that can generate a synthetic image as close as possible to given input and to use resulting AAM parameters for interpretation [4]. There has been developed several AAM search strategies that are precise enough to detect facial features. However, in our case we decided to use manually placed landmark points for shape definition, because AAM search can sometimes result in false detections which would decrease final emotion and emotion intensity level classification accuracy.

An AAM was employed for extraction of facial expression features of three types: shape parameters vector $\mathbf{x}$, texture parameters vector $\mathbf{g}$ and combined parameters vector $\mathbf{c}$. The amount of detail, or variance Ψ, contained in AAM model can be controlled by changing the number of modes of variation (number of eigenvectors) contained in $\mathbf{P}_s$, $\mathbf{P}_g$, $\mathbf{P}_c$. This number of variation modes determines the size of $\mathbf{x}$, $\mathbf{g}$ and $\mathbf{c}$.

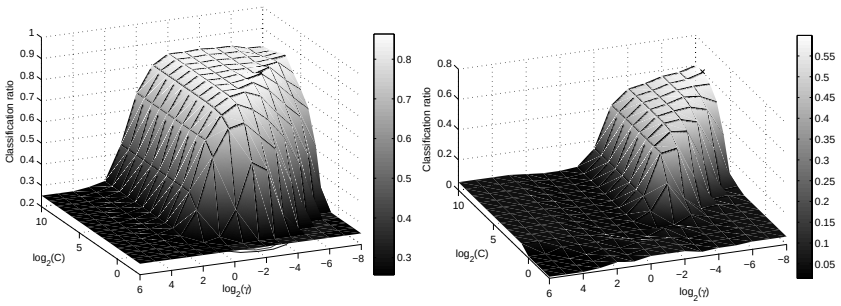
3 Support Vector Machine classifier

SVM classifiers with Gaussian RBF kernels were used for the task of emotion and emotion level classification. SVM classifiers are well known for their good generalization properties even in cases of high-dimensional nonlinear separable classification tasks. The RBF kernel was chosen because it has fewer adjustable parameters than any other commonly used kernel and has less numerical difficulties [3].

SVM classifiers can work well only when optimal values for its parameters are set. SVM classifier with Gaussian RBF kernel has two parameters: 1) the SVM regularization constant $C > 0$ and 2) γ parameter related to Gaussian RBF kernel function $k(\mathbf{x}, \mathbf{y}) = exp(\frac{-\|\mathbf{x}-\mathbf{y}\|^2}{\gamma})$. V-fold cross-validation and Grid-search algorithm can be utilized for this purpose. In v-fold cross-validation, the training set is first split into v subsets of equal size. Sequentially the classifier with defined learning parameters is trained v times, where in the i-th iteration ($i = 1, \ldots$,v) it is trained on all subsets except the i-th one. The classification error E_i is computed for the i-th subset. This procedure calculates v values of classification error where the average value of them $E = \frac{1}{v}\sum_{i=1}^{v} E_i$ is a rather good estimate of the classifier generalization error. The precision of classifier generalization error estimation

can be improved when several v-fold cross-validations (with different initial training set split) is performed and their results are averaged. The same approach can help to improve training accuracy when training data has small number of samples.

Grid search basically tries to estimate the generalization error of classifier using k ($k = mn$) different pairs of (C, γ) ($\{(C,\gamma)|C \in \{C_1, C_2, \ldots, C_m\}; \gamma \in \{\gamma_1, \gamma_2, \ldots, \gamma_n\}\}$). The pair with lowest generalization error is selected as an optimal. It was found [3] that trying exponentially growing sequences of C and γ is a practical method to identify good parameter values.



(a) (b)

Figure 1: Visualization of v-fold cross-validation results. Grid-search approach ($C = \{2^{-3}, 2^{-2}, \ldots, 2^{10}\}$, $\gamma = \{2^{-8}, 2^{-7}, \ldots, 2^{6}\}$) was used to find the optimal values (the black crosses) for the RBF SVM classifier parameters. (a) Emotion classifier, (b) Happiness intensity levels classifier

4 Experiments

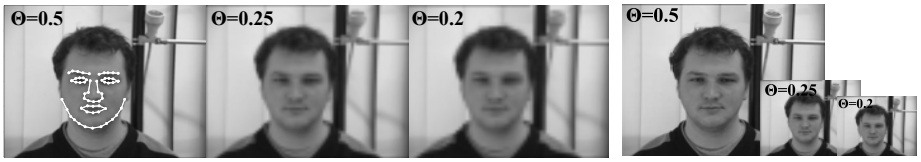
Facial emotion images were obtained from FG-NET facial emotion database [14] containing face image sequences showing a number of Caucasian subjects performing the six different basic emotions defined by Eckman [6]. One of the underlying paradigms of this database is to let the observed people react as naturally as possible (real emotions were elicited by showing video clips or still images after a short introduction phase). We decided to exclude emotions of sadness and fear because according to our observations they do not reach the same level of intensity and are not as distinctive as the other basic emotions. Images of selected emotions comprising anger, happiness, disgust and surprise (3 emotion levels for each emotion) plus neutral expression images of 12 individuals were selected from the FG-NET database and used during the analysis (see Figure 2(a)). The particular emotion intensity level images were selected from the image sequences on the basis of subjective selection. Additionally, all 192 faces were manually labeled with 58 landmark points defining face shapes. The placement of landmark points can be seen in Figure 2(b) (first image from left).

One of our aims was to analyze how the image quality affects the recognition accuracy. For this purpose we used source images (size 320×240) to generate images with reduced size. Size reduction factors $\Theta = \{0.5, 0.25, 0.2\}$ and bilinear interpolation method were used for generating of images with reduced size (see Figure 2(c)). In the case of psycho-

logical experiment the images were too small so they were enlarged to size 800×600 (see Figure 2(b)).

Images of 8 subjects (4 males, 4 females) served as training samples for AAM and SVM classifiers. The remaining images (2 males, 2 females) form the testing set for SVM classifier as well as for psychological experiment.

(a)



(b) (c)

Figure 2: (a) Examples of expression emotion images. 4 individuals performing emotion of disgust, surprise, anger and happiness with 4 emotion level intensities (neutral face expression stands for zero emotion intensity level). Examples of image quality degradation for psychological experiment (b) and size reduction for SVM training (c). Landmark points defining face shape can be seen in the most left image of (b).

.

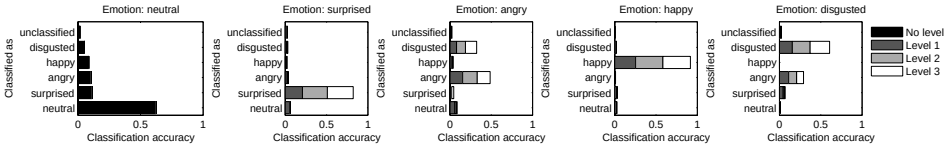
4.1 Psychological Experiment

Psychological experiment which should evaluate the ability of human respondents to recognize facial emotions and emotion intensity levels was prepared. Each trial of experiment consists of three parts - slides. In the first slide, an unmasked facial emotion image is presented and participant has 5 sec. to classify one of 5 emotions using the computer mouse to make the choice. Another slide with emotion intensity levels choices is shown for the same image. Again, the respondent has 5 sec. to choose the right choice. If the participant does not meet the time limit a "nothing" string value is saved as the response in both cases. Furthermore, if participant classify emotion in the first slide as neutral then the chosen level response in the second slide is not recorded. The last slide has no visual content and lasts for 100 ms, it serves as separator between trials.

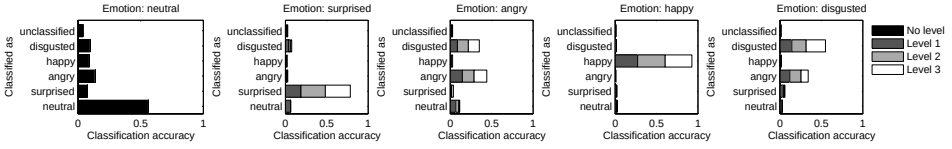
In each experimental session, participants perform 192 trials shown in random order, which followed a 4 individual $\times$ 4 emotions $\times$ 4 levels (neutral expression which serves as zero emotion intensity level is added to mentioned three emotion intensity levels) $\times$ 3 different image quality design. Twenty-four volunteers (12 males, 12 females) with normal or corrected-to-normal vision participated in this study. They had no prior knowledge about the subjects. They were instructed in detail and images of 2 individuals (1 male, 1 female, all emotions and emotion levels, size reduction factor $\Theta = 0.5$) with correct

information about their emotional state were shown to them before the experiment.

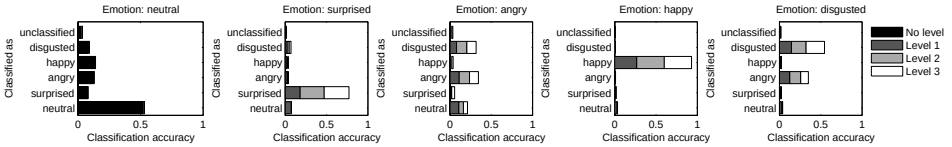
It can be seen from the analysis of results shown in Figure 3 that positive emotions comprising happiness and surprise are recognized with very good accuracy, and almost independently on image reduction. Negative emotions including anger and disgust are often mistaken and also neutral expression is often misclassified. It can be also seen how the accuracy decrease witch increasing Θ. The analysis of results for emotion intensity level classification can be found in Figure 4. It can be summarized that neighboring levels were often mistaken and that highest level was chosen the fewest times. The final classification accuracy can be found in Table 1.



(a) $\Theta = 0.5$



(b) $\Theta = 0.25$



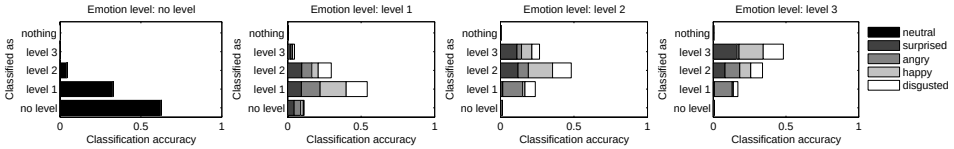
(c) $\Theta = 0.2$

Figure 3: Analysis of emotion classification results retrieved from psychological experiment. The histograms show what were the respondent's choices in comparison with the original emotion labels and bar colors shows how the original emotion levels affected the respondents choices.

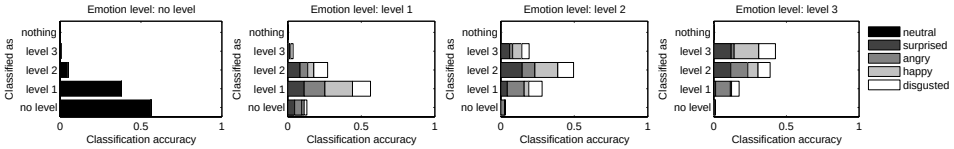
4.2 SVM Classifications

Thirty-six types (3 sizes of source facial images ($\Theta = \{0.5, 0.25, 0.2\}$) $\times$ 3 AAM parameter vector types (shape, texture, combined) $\times$ 4 values for variance retained in AAM models ($\Psi = \{0.93, 0.95, 0.97, 0.99\}$) of training data were tested for SVM classifier training. Grid-search approach ($C = \{2^{-3}, 2^{-2}, \ldots, 2^{10}\}$, $\gamma = \{2^{-8}, 2^{-7}, \ldots, 2^{6}\}$) was used to find the optimal values for the RBF SVM classifier parameters. Examples of grid-search classification results can be seen in Figure 1(a) (emotion classifier, results averaged over 4 tenfold cross validations (combined parameters vectors, $\Psi = 0.93$, $\Theta = 0.2$)) and 1(b) (happiness intensity level classifier, results averaged over 4 sixfold cross validations (shape parameters vectors, $\Psi = 0.93$, $\Theta = 0.2$)).

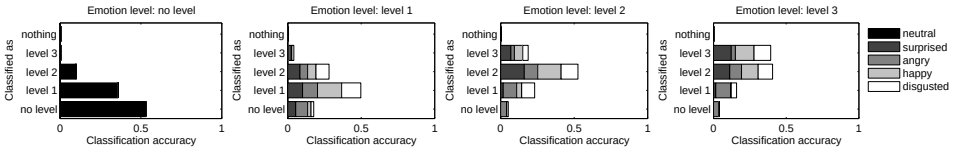
Five classifiers (1 classifier of emotions (128 training samples), 4 emotion intensity



(a) $\Theta = 0.5$



(b) $\Theta = 0.25$



(c) $\Theta = 0.2$

Figure 4: Analysis of emotion intensity level classification results retrieved from psychological experiment.

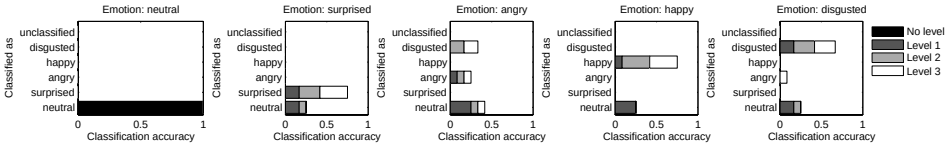
	SVM		**Respond.**	
Θ	**Emotions**	**Levels**	**Emotions**	**Levels**
0.5	**70.31%**	50.00%	68.75%	**53.26%**
0.25	**65.63%**	45.31%	64.71%	**51.11%**
0.2	**64.06%**	43.75%	61.85%	**48.70%**

Table 1: Final classification accuracy.

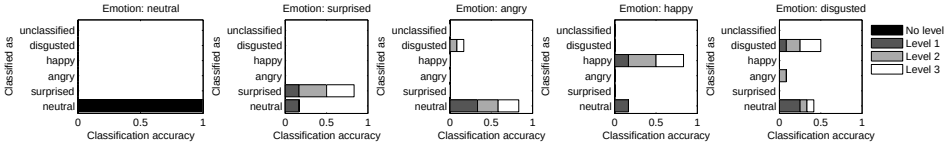
levels classifiers (32 training samples)) were trained using data type and SVM parameters that achieved best classification accuracy. Analysis of SVM classification results can be seen in Figure 5 and 6. The neutral emotion and "no level" choice are preferred by the classifiers because neutral images are presented with the highest number of samples in the training set. This can be seen especially in the case of negative emotions that are not as visually distinctive as positive emotions and in the case of first emotion intensity level.

5 Discussion and Conclusions

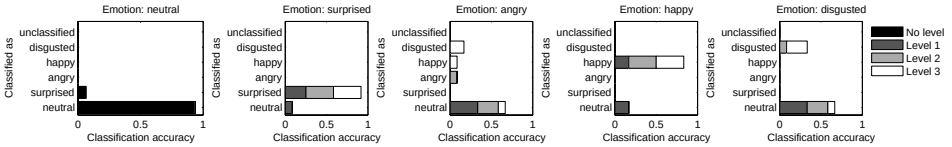
It can be seen in the Table 1 that SVM classifier outperforms the average classification accuracy of human respondents in the case of emotion classification for all image sizes. However, the SVM failed in the second case of emotion level intensity classification. This failure is probably caused by the small training set size. Moreover, the decrease of image sizes reduces the classification accuracy of both SVM classifier and the average accuracy



(a) $\Theta = 0.5$

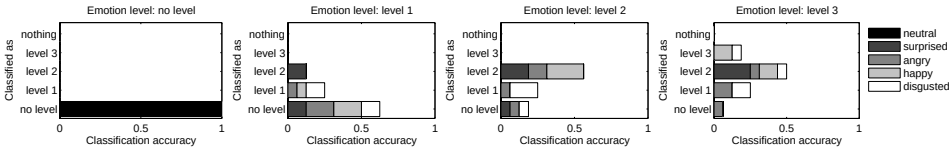


(b) $\Theta = 0.25$

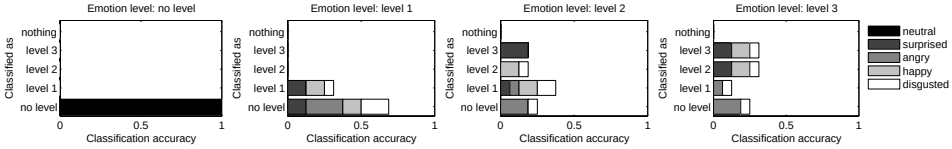


(c) $\Theta = 0.2$

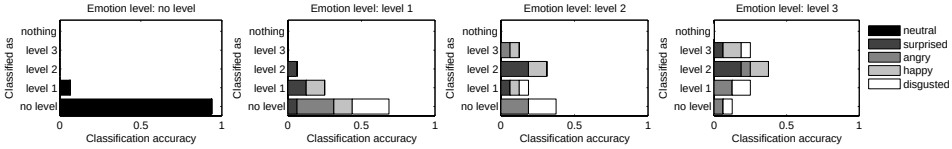
Figure 5: SVM classifier emotion classification analysis.



(a) $\Theta = 0.5$



(b) $\Theta = 0.25$



(c) $\Theta = 0.2$

Figure 6: SVM classifier emotion intensity level classification analysis

.

of human respondents. This is caused by the loss of visual facial details that can be used for differentiation between emotions especially in the case of lower intensity levels.

				SVM				Respondents			
Emotion	*N*	*P*	Scale	*TP*	*FP*	*Acc*	*Prc*	*TP*	*FP*	*Acc*	*Prc*
neutral	48	16	0.5	16	14	0.78	0.53	10	2	**0.87**	**0.82**
			0.25	16	19	0.70	0.46	9	3	**0.85**	**0.77**
			0.2	15	19	0.69	0.44	9	4	**0.82**	**0.66**
surprised	52	12	0.5	9	0	**0.95**	**1.00**	10	4	0.91	0.74
			0.25	10	0	**0.97**	**1.00**	9	3	0.92	0.78
			0.2	11	1	**0.97**	**0.92**	9	2	0.92	0.80
angry	52	12	0.5	3	1	**0.84**	**0.75**	6	6	0.81	0.50
			0.25	1	0	**0.83**	**1.00**	5	7	0.79	0.44
			0.2	1	0	**0.83**	**1.00**	4	7	0.77	0.38
happy	52	12	0.5	9	0	**0.95**	**1.00**	11	2	0.95	0.84
			0.25	10	0	**0.97**	**1.00**	11	2	0.95	0.84
			0.2	10	1	**0.95**	**0.91**	11	3	0.93	0.77
disgusted	52	12	0.5	8	4	**0.88**	**0.67**	7	5	0.84	0.58
			0.25	6	2	**0.88**	**0.75**	7	7	0.81	0.49
			0.2	4	2	**0.84**	**0.67**	7	6	0.82	0.52
Level 1	32	16	0.5	4	8	0.58	0.33	9	7	**0.71**	**0.57**
			0.25	5	8	0.60	0.38	9	7	**0.70**	**0.55**
			0.2	4	7	0.60	0.36	8	6	**0.70**	**0.56**
Level 2	32	16	0.5	9	10	**0.65**	**0.47**	8	10	0.62	0.43
			0.25	3	5	**0.63**	**0.38**	8	11	0.61	0.43
			0.2	5	7	**0.63**	0.42	8	11	0.61	**0.43**
Level 3	32	16	0.5	3	0	**0.73**	**1.00**	8	5	0.72	0.61
			0.25	5	3	0.71	0.63	7	4	**0.73**	**0.65**
			0.2	4	2	0.71	**0.67**	6	4	**0.72**	0.63

Table 2: Analysis of ROC space characteristics. *Acc* - Accuracy, *Prc* - Precision, *N*/*P* - counts of negative/positive samples in testing set, *TP*/*FP* - counts of true positives / false positives classifications. $Acc = (TP + (N - FP))/(P + N)$; $Prc = (TP)/(TP + FP)$

The receiver operating characteristic space analysis for all source image sizes can be found in Table 2. In this case we treat the classification results as results retrieved from binary classifiers for individual emotions or emotion levels. We utilized Accuracy and Precision ROC space characteristics to compare average classification results of human respondents and classification results of SVM. The results highlight the variability innate in the human participant's responses to emotional cues. The emotional levels of image set were labeled by one person, who placed its interpretation on emotional content. It is therefore inaccurate to say that people are worse or better than the SVM in recognizing emotional cues in the case of emotional levels. It is better to say that SVM is more consistent than people in this study at labeling the gross class categories, with the exception of neutral emotion. Conversely, people were more consistent with emotional level identification than the SVM.

We have proposed an alternate and general way of classifier evaluation of emotional cues. This kind of evaluation is independent of testing data set structure and complexity. It can help to get classification quality quantifiers that can be compared with other methods

in more a human-centred and reliable way.

References

[1] J. N. Bassili. Facial motion in the perception of faces and of emotional expression. *Journal of Experimental Psychology Human Perception and Performance*, 4:373–379, 1978.

[2] M. Beszedes and M. Oravec. Adaboost algorithm used for skin color detection. *29th International Conference Telecommunications and Signal Processing TSP-2006 Brno, Czech Republic*, pages 96–99, Sept 2006.

[3] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[4] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[5] G.J. Edwards, T.F. Cootes, and C.J. Taylor. Face recognition using active appearance models. *Proceedings of the European Conference on Computer Vision*, 2:581–695, 1998.

[6] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17, 1971.

[7] B. Fasel and J. Luettin. Automatic facial expression analysis: A survey. *Pattern Recognition*, 36(1):259–275, 2003.

[8] J. Liebelt, J. Xiao, and J. Yang. Robust aam fitting by fusion of images and disparity data. *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Volume 2*, pages 2483–2490, 2006.

[9] M.J. Lyons, J. Budynek, and S. Akamatsu. Automatic classification of single facial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12):1357–1362, 1999.

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

[11] M. Pantic and L.J.M. Rothkrantz. Automatic analysis of facial expressions: the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 2000.

[12] J.A. Russell. Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies. *Psychological Bulletin*, 115(1):102–141, 1994.

[13] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *Proc. CVPR*, 1:511–518, 2001.

[14] F. Wallhoff. Facial expressions and emotion database. `http://www.mmk.ei.tum.de/~waf/fgnet/feedtum.html`, 2006.