Cascaded Classification of Gender and Facial Expression using Active Appearance Models

Yunus Saatci and Christopher Town
University of Cambridge Computer Laboratory
15 JJ Thomson Avenue, Cambridge, CB3 0FD, UK
{ys267, cpt23}@cam.ac.uk

Abstract

This paper presents an approach to recognising the gender and expression of face images by means of Active Appearance Models (AAM). Features extracted by a trained AAM are used to construct Support Vector Machine (SVM) classifiers for 4 elementary emotional states (happy, angry, sad, neutral). These classifiers are arranged into a cascade structure in order to optimise overall recognition performance. Furthermore, it is shown how performance can be further improved by first classifying the gender of the face images using an SVM trained in a similar manner. Both gender-specific expression classification and expression-specific gender classification cascades are considered, with the former yielding better recognition performance. We conclude that there are gender-specific differences in the appearance of facial expressions that can be exploited for automated recognition, and that cascades are an efficient and effective way of performing multi-class recognition of facial expressions.

1. Introduction

The recognition of facial attributes such as identity, gender, expression, ethnicity, and age has attracted much attention in computer vision. An idea that has been less explored is the interdependency of recognition performance in these tasks based on gender. Studies in psychophysics have shown that such linkages exist, with women generally performing better at expression recognition (although men are better at detecting signs of anger [15]). At the same time, male faces appear to exhibit greater variability that may aid in identification. For example, [10] notes that facial attractiveness for men is inversely related to recognition accuracy, but not for women.

Some computer vision researchers have also noted difference in face recognition performance with respect to the gender of the face images [12]. In evolutionary terms, one might explain such differences as a results of sexual selection and cultural influences [15] that have shaped male and female physical appearance and performance on visual social communication tasks such as body language and empathy. In many societies women are expected to be more emotionally extrovert while men are required to maintain a more guarded "poker face".

In this paper, we first describe the use of AAMs as a feature extraction mechanism in two common facial interpretation tasks: expression classification and gender classification. In Section 4 we describe the architecture and performance of an expression recognition system that uses the AAM framework for feature extraction and Support Vector Machines (SVMs) for classification. In Section 5, we delineate a similar system that performs gender classification. By iteratively optimising accuracy over a test set, we construct an optimal cascade consisting of binary SVM classifiers for a set of 4 basic expressions. In Section 6 we explain how the cascade can be further extended by combining gender and expression classification. We achieve best results using a tree structure consisting of two expression classification cascades that were selectively trained on male and female images respectively as determined by an initial gender classifier.

2. Related Work

2.1. Facial Expression Recognition

In the recent past, many different approaches for facial expression classification in static images have been evaluated. Pantic and Rothkrantz [11] survey the large variety of such techniques and find that all of them consist of three major steps: face detection, a mechanism for extracting facial expression information, and a mechanism to classify the information extracted according to some pre-defined set of categories. In order to implement the second step, past

algorithms have used *feature-based*, *template-based* or *hybrid* face representations. Feature-based face representations model the face as a set of facial points, whereas in template-based representations the face is represented as a whole unit. There is also a certain amount of controversy as to what the classes should be in such a classification system. This is reflected by the fact that certain expression recognisers classify the encountered expression as a particular set of facial actions (such as "raised eyebrows" and "open mouth") whereas others classify basic emotions, as described by Ekman [3]. These basic emotions include "joy", "sadness", "anger", "surprise", "fear" and "disgust" (along with "neutral", which indicates a lack of emotion).

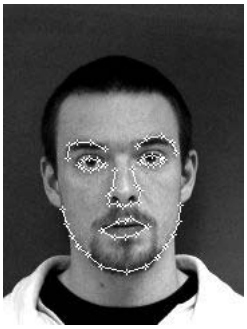

Figure 1. An instance of the dataset used to construct the AAM.

2.2. Gender Classification

Although gender classification has attracted the interest of many cognitive psychologists, the number of attempts made at automating the process have been fewer in comparison. The first attempt was made by Gollomb et. al. ([5]) who trained a multi-layer neural network, SEXNET, to classify gender in 90 image samples of men and women. Brunelli and Poggio [1] followed a feature-based approach where two competing Radial Basis Functions (RBFs) (one for male and one for female) were trained on the geometric relationships between facial features. Moghaddam and Yang [9] proposed a non-linear SVM for gender classification using the FERET database where the feature vectors for the SVMs were given by the greyscale values of "thumbnail" face images. They quote an error rate of 3.4%, which seems to be the best result in the open literature.

3. Active Appearance Models

In order to construct an AAM [2] for full-frontal faces, it is necessary to have a labelled training set in which each image is accompanied with data specifying the coordinates of landmark points (usually at least 20) (see Figure 1). The appearance model is then obtained by constructing a shape model using the coordinate data (which can be viewed as a set of *shape vectors*) and a texture model using both the image data and the coordinate data. The shape model is built by aligning all of the shape vectors to a common coordinate frame and performing Principal Component Analysis (PCA) on these. The shape model is then controlled by $\mathbf{b}_s$, each shape generated by the model calculated using, $\mathbf{x} = \overline{\mathbf{x}} + \mathbf{P}_s\mathbf{b}_s$; where $\mathbf{P}_s$ contains the eigenvectors of the sample covariance matrix and $\overline{\mathbf{x}}$ is the mean of the aligned shape vectors. To construct a model of gray-levels, each training image is warped so that the control points match the mean shape $\overline{\mathbf{x}}$, using Delaunay triangulation to calculate the warp parameters. The texture data within the region of the face now bounded by the mean shape is sampled to form a texture vector $\mathbf{g}_{raw}$. To minimize the effects of global lighting variations, histogram equalization is applied to $\mathbf{g}_{raw}$ to form a normalised texture sample $\mathbf{g}_i$ which can be used in statistical analysis. In a similar manner to the shape model, the texture model is constructed from $\{\mathbf{g}_i\}$ using PCA. In order to obtain further dimensionality reduction, the process of building an AAM also involves applying a further PCA to the shape and texture data. This results in an "appearance model" that controls both shape and texture. A variant of the Simulated Annealing algorithm is used to learn how to update model parameters from pixel errors during search. This information is combined with the appearance model to form the "Active Appearance Model". By modelling the complete texture and shape of faces, along with incorporating an efficient model optimisation mechanism, the

4. Expression Recognition using AAMs

We chose to classify four of the basic emotions ("happy", "angry", "sad" and "neutral") outlined by Ekman [3]. In order to initialise the AAM search, we use an implementation of the Viola-Jones face detector [14].

4.1. Feature extraction using AAMs

As described in Section 3, in order to train an AAM which would be useful for our expression classification system, it was necessary to have an annotated dataset of full-frontal face images in which subjects display each of the four expressions genuinely. It was also important that the faces in the training set had different appearances as determined by gender, race and the existence of facial hair and/or glasses, since this would improve the generalisation performance of the final recognition system. With these considerations in mind, we opted to work on a dataset of 1,135 samples drawn from the Purdue AR dataset [8], the IMM dataset by the Denmark Technical University [13] and the

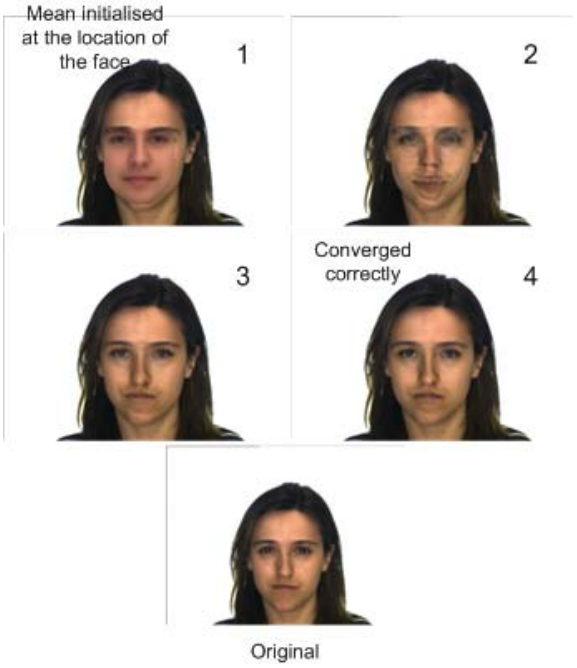

Figure 2. An example of successful AAM convergence.

FEEDTUM dataset of the FG-NET consortium [1]. Only 74 of the 1,135 samples were found to be annotated sufficiently well for our purposes (with 58 landmark points), so we had to employ a bootstrapping procedure (see Table 1) to expand the size of the training set in order for the AAM to have an acceptable correct convergence rate. As a result of

```
initialise the current training set;
repeat
  train AAM on current training set;
  test AAM on the rest of the dataset;
  add cases that have correctly converged
  to the current training set;
until correct convergence rate is acceptable
```

Table 1. Bootstrapping algorithm.

this procedure, we managed to expand the training set to one with 262 fully annotated samples. This training set resulted in an AAM which had a good enough correct convergence (search) rate of 81%. In total, the AAM converged correctly on 809 image samples. Thus, a total of 809 feature vectors were derived for use in training and testing the classifiers in the final stage. See Figure 2 for an illustration of the AAM search procedure. It is also worth noting that the final AAM had 60 control parameters and hence 60 dimensional feature vectors.

[1] Face and Gesture Recognition Network

4.2. Classification

A number of schemes for multi-class classification have have been proposed [4]. Hierarchical decomposition of a feature space by means of a tree or cascade structure (e.g. [14]) has been shown to be highly effective, and allows binary classifiers such as SVMs to be applied to complex problems [6].

Using a random subset of 60% of the feature vectors, we trained a *cascade* of SVM classifiers. Each binary SVM classifier in the cascade was trained to act as an *expression detector*, outputting a positive response if its expression is present and a negative response otherwise. So, for example, a binary SVM trained as a "happy detector" would classify between expressions which are *happy* and *not happy*. We trained SVMs with linear, polynomial and RBF (Radial Basis Function) kernels in order to compare performance at the testing stage. For SVM kernels with parameters (i.e. the polynomial and RBF kernels) we needed to perform a search for the best parameter value over a restricted subspace of the parameter space. For the polynomial kernel function SVM, the recommended subspace to use was all integer degrees p ranging from 1 to the number of degrees-of-freedom within the feature vectors (in this case 60 - recall that the equation for this kernel is $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + 1)^p$). For the RBF kernel with parameter γ, the subspace of search was taken as $\gamma = 2^{-15}, 2^{-14}, \ldots, 2^2, 2^3$. Thus, for each binary classifier, we tested 1 linear, 60 polynomial and 19 RBF SVMs.

We then built a cascade structure by iteratively adding the classifier that:

(a) has the highest accuracy (as measured by the jointly optimal detection and false positive rate on the ROC curve, i.e. the "top-most left-most" point on the curve) for the given expression,

(b) has the lowest false positive rate out of the remaining expression classifiers satisfying condition (a).

	H	**S**	**A**	**N**	**U**
H	94.4%	0%	0%	1.12%	4.49%
S	0%	70.5%	14.8%	9.84%	4.92%
A	1.64%	4.92%	77.1%	8.14%	8.20%
N	2.60%	13.0%	19.5%	63.6%	1.30%

Table 2. Confusion matrix for the SVM cascade classifying AAM feature vectors.

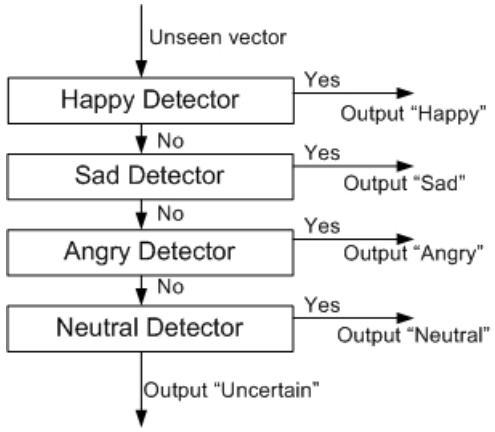

Figure 3. Expressional classification cascade.

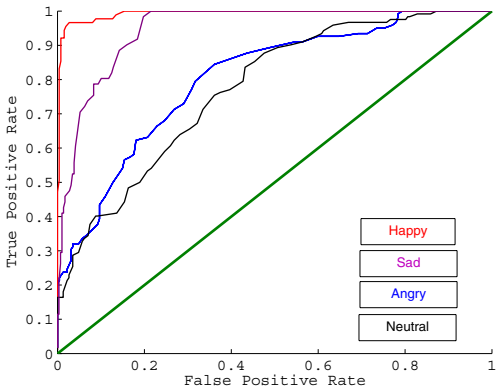

Figure 4. ROC curves for the binary SVM classifiers.

In our case, the cascade had the structure shown in Figure 3. The ROC curves for the binary SVM classifiers used in the final cascade are shown in Figure 4. The confusion matrix for the overall cascade is shown in Table 2. In all confusion matrices, "H", "S", "A", "N" and "U" stand for "Happy", "Sad", "Angry", "Neutral" and "Unrecognised" respectively. The overall accuracy of the cascade was calculated as 76.4%.

5. Gender Classification using AAMs

After obtaining AAM feature vectors from the full-frontal face image dataset, gender classification involved constructing a training set by using 60% of the feature vectors (randomly chosen) to train a family of SVMs (just as in Section 4.2). The SVM which gave the highest accuracy whilst being tested on the remainder (40%) of the dataset was chosen to be the most superior gender classifier. The optimal classifier was found to have a staggering accuracy of 97.6% and a false positive rate of 0.735%. The area under the ROC curve, as shown in Figure 5, was close to its ideal value: 0.986. In the ROC curve, the positive class was arbitrarily chosen to correspond to the male class.

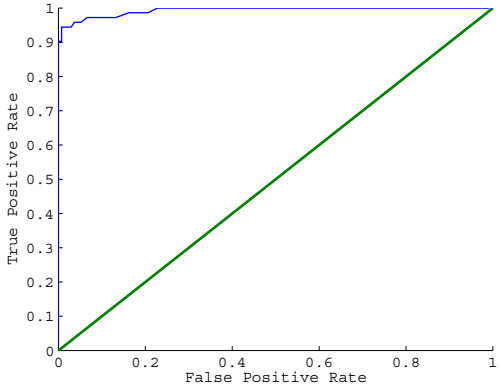

Figure 5. ROC curve for the optimal gender classifier.

6. Combining Expression and Gender Classification

Research in neurobiology seems to indicate that biological visual systems are structured in a hierarchical fashion with mutual feedback occuring between different levels in the hierarchy [7]. The existence of cells in the visual cortex which can be triggered by stimuli of ever-increasing complexity [2] and the massive feedback loop from the visual cortex to the lateral geniculate nucleus (which is at a lower level than the cortex) clearly support this hypothesis. Therefore, it is likely that the human brain performs the functions of expression and gender recognition at different levels in its complex hierarchy and that there exist feedback paths between these two levels. With this in mind, we decided to test whether the performance of our expression classification system could be improved if we performed gender classification in advance and then fed the gender predictions to gender-dependent expression classifiers. Likewise, we decided to test whether the performance of our gender classification system could be improved if we performed expression classification in advance and then fed the expression

[2] Recent research even suggests that we have singleton top-level cells that respond only when we see people we recognise.

predictions to expression-dependent gender classifiers. We describe both these systems below.

6.1. Expression Classification using Gender Recognition

The architecture of the system in which gender recognition is used to inform gender-specific expression classifiers is shown in Figure 6.

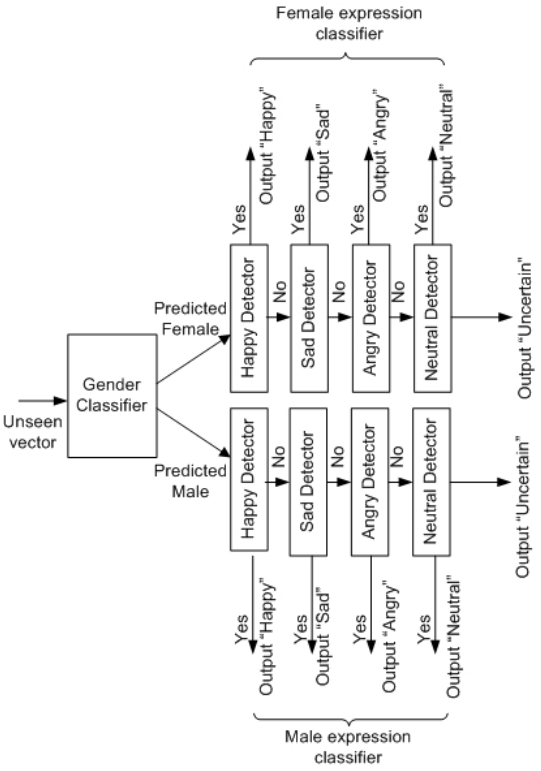

Figure 6. Expression Classification using Gender Recognition.

The gender classifier shown in Figure 6 is identical to the classifier used in Section 5. The gender-specific expression classification cascades were trained as in Section 4; however, the training data for each binary SVM was drawn strictly from either male or female samples, depending on which gender-specific cascade the binary SVM lived in. Note that we may expect the gender-specific expression classifiers to perform in a superior fashion since fixing the gender variable has the consequence of reducing intra-class variability. However, it is also worth noting that training each binary SVM on either male or female samples reduced the size of the training sets for each, and that this could diminish the positive effect of reduced intra-class variability.

	H	S	A	N	U
H	92.2%	0%	1.82%	3.64%	2.38%
S	0%	75.6%	13.1%	6.72%	4.64%
A	0.77%	3.18%	70.7%	12.3%	13.1%
N	0%	1.67%	4.86%	81.0%	12.5%

Table 3. Confusion matrix for expression classification using gender recognition.

The overall cascade was tested on all samples not used to train the binary SVM emotion detectors and the results are shown in the confusion matrix in Table 3. The overall accuracy was found to be 79.9%, which shows an improvement over not using gender classification at all. We expect that the improvement would have been more visible if the binary SVM classifiers in the gender-specific expression recognisers were trained with datasets of a similar size to those in Section 4.

	H	S	A	N	U
H	97.1%	0%	0%	0%	2.94%
S	0%	80.7%	10.1%	5.32%	3.88%
A	0.98%	1.34%	77.7%	11.3%	8.73%
N	0%	1.04%	3.44%	88.2%	7.32%

Table 4. Confusion matrix for expression classification in females using gender recognition.

	H	S	A	N	U
H	87.3%	0%	3.64%	7.28%	1.82%
S	0%	70.5%	16.1%	8.12%	5.40%
A	0.56%	5.02%	63.7%	13.3%	17.5%
N	0%	2.30%	6.28%	73.7%	17.7%

Table 5. Confusion matrix for expression classification in males using gender recognition.

It is also interesting to note, by observing confusion matrices in Tables 4 and 5, that expressions in female faces tend to be classified more successfully in comparison to those in male faces!

6.2. Gender Classification using Expression Recognition

Given that expression classification can be improved using gender recognition, it may be the case that inverting this hierarchical relationship might cause an improvement in gender classification. The system constructed to evaluate this hypothesis using the AAM feature-vector set

is depicted in Figure 7. The cascaded expression classifier shown is identical to the one used in Section 4. Each expression-specific gender classifier was trained using AAM feature vectors drawn from the relevant expression class. For feature-vectors which would not be recognised by the expression recogniser, it was necessary to use the "general", expression-independent gender classifier as built in Section 5. Note that, as in the previous Section, each expression-specific gender classifier benefits from the lowered intra-class variation (due to the expression being fixed) yet is trained on smaller training sets.

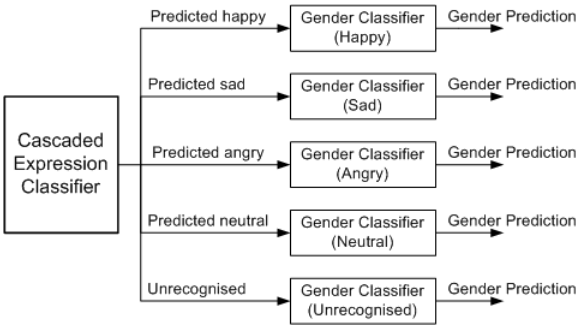

Figure 7. Gender Classification using Expression Recognition.

	ACC	TP	FP
Happy	90.7%	86.4%	4.76%
Sad	95.7%	94.6%	2.21%
Angry	97.9%	96.7%	1.32%
Neutral	95.5%	96.0%	4.95%
Unclassified	94.3%	92.1%	2.34%
Average	94.8%	93.2%	3.12%

Table 6. Confusion matrix for gender classification using expression recognition.

The confusion matrix shown in Table 6 illustrates the results obtained. ACC, TP, FP stand for Accuracy, True Positive Rate and False Positive Rate respectively. Although the overall results are poorer than those obtained in Section 5, this was probably due to the (approx.) 4-fold decrease in the size of the training sets used to train each expression-specific gender classifier.

7. Summary and Conclusions

We have described the architecture and performance of an expression and gender recognition system that uses AAMs for feature extraction and SVMs for classification. By iteratively optimising accuracy over a test set, we constructed an optimal cascade consisting of binary SVM classifiers for a set of 4 basic expressions. The cascade performs significantly better at recognition and disambiguation than other classification combination schemes such as maximum margin. We further show that performance can be improved further by combining gender and expression classification. Best results were obtained using a tree structure consisting of two expression classification cascades that were selectively trained on male and female images respectively as determined by an initial gender classifier.

References

[1] R. Brunelli and T. Poggio. Hyperbf networks for gender classification. In *DARPA Image Understanding Workshop*, pages 311–314, 1992.

[2] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *Lecture Notes in Computer Science*, 1407:484–502, 1998.

[3] P. Ekman. *Emotion in the Human Face*. Cambridge University Press, 1982.

[4] J. Ghosh. Multiclassifier systems- back to the future. In *Proc. 3d Int. Workshop on Multiple Classifier Systems*, 2002.

[5] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski. Sexnet: A neural network identifies sex from human faces. *Advances in Neural Information Processing Systems*, pages 572–577, 1991.

[6] H. Graf, E. Cosatto, L. Bottou, I. Durdanovic, and V. Vapnik. Parallel support vector machines: The cascade svm. In *NIPS*, 2004.

[7] D. R. Hofstadter. *Gödel, Escher, Bach: an Eternal Golden Braid*. Penguin Books, 1979.

[8] A. Martinez and R. Benavente. The AR face database. Technical Report 24, CVC Technical Report, June 1998.

[9] B. Moghaddam and M.-H. Yang. Sex with support vector machines. In *NIPS*, pages 960–966, 2000.

[10] A. OToole, K. Deffenbacher, D. Valentin, K. McKee, D. Huff, and H. Abdi. The perception of face gender: the role of stimulus structure in recognition and classification. *Memory and Cognition*, 26(1), 1998.

[11] M. Pantic and L. J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1424–1445, 2000.

[12] P. Phillips, P. Grother, R. Micheals, D. Blackburn, E. Tabassi, and J. Bone. FRVT 2002: Overview and summary, 2002.

[13] M. B. Stegmann, B. K. Ersbøll, and R. Larsen. FAME – a flexible appearance modelling environment. *IEEE Trans. on Medical Imaging*, 22(10):1319–1331, 2003.

[14] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *ICCV*, 2001.

[15] S. Widen. Gender and preschoolers' perception of emotion. *Merrill-Palmer Quarterly*, 48(3), 2002.