

Gender Classification Using Mixture of Experts from Low Resolution Facial Images

Yomna Safaa El-Din¹, Mohamed N. Moustafa², and Hani Mahdi¹

¹ Department of Computer and Systems Engineering,
Ain Shams University,
Cairo, Egypt

{yomna.safaa-eldin,hani.mahdi}@eng.asu.edu.eg

² Department of Computer Science and Engineering,
American University in Cairo,
New Cairo, Egypt

moustafa@ieee.org, m.moustafa@aucegypt.edu

Abstract. In this study, we propose a novel two-stages mixture of experts scheme estimating gender from facial images. The first stage combines a couple of complementary gender classifiers with a third arbiter in case of decision discrepancy. Experimentally, we have verified the common thinking that one appearance-based (Haar-features cascade) classifier with another shape-based (landmarks positions metrology with SVM) classifier form a complementary couple. Subsequently, the second stage in our scheme is a Bayesian framework that is activated only when the arbiter cannot take a confident decision. We demonstrate that the proposed scheme is capable of classifying gender reliably from faces as small as 16x16 thumbnails on benchmark databases, achieving 95% gender recognition on FERET database, and 91.5% on the Labeled Faces in the Wild dataset.

Keywords: gender classification, committee machines, Bayes, resolution.

1 Introduction

Many human activities and machine applications depend on accurate gender recognition. It can be used as a prior step to face recognition and verification [1,2]. Gender discrimination also helps in the indexing and retrieval of images and videos [3].

Automatic gender classification has been widely investigated in literature. Most studies have used 2D face images for classification, which can be done using either appearance-based or shape-based methods. Appearance-based approaches use the cropped, resized, and illumination normalized texture of the face (or portions of it) as a classification attribute, while the shape-based approaches extract a set of discriminative face shape features and uses them for the classification process.

It is expected that the classification accuracy can be improved by combining more than one classifier [4], especially when each method relies on different input features extracted from the face.

In this work we compare several state-of-the-art classification techniques on different features and image resolution. Then we propose a Mixture of Experts (MoE) technique based on Naive Bayes theorem for merging some of these well known classifiers to achieve boosted performance.

The rest of the paper is organized as follows; Section 2 reviews previous related work, Section 3 describes the individual classification methods used along with their input feature types, and Section 4 introduces our proposed method for merging these classifiers. Section 5 states the databases we used, then explains and discusses the experiments and results. In the final section, we conclude our work.

2 Related Work

Mäkinen et al. [4] presented an overview on the topic of gender classification from face images. They experimented on FERET database [5] and WWW [4] (another dataset containing images they randomly collected from the web). They compared six state-of-the-art gender classification algorithms, none of which is shape-based. They used different face normalizations and alignments, and they introduced combined results of these classifiers.

Another comparative study was presented by Calfa et al. in [6], giving special attention to linear techniques and their relations, due to their simplicity and low computational requirements. Their work proves that, with a linear feature selection, Linear Discriminant Analysis on the linearly selected set of features achieves results comparable to the best gender classifiers based on Support Vector Machines with Radial Basis Function kernel (SVM+RBF) [7] and Boosting.

Shan [8] investigated gender classification on real-life faces acquired in unconstrained conditions. Boosted Local Binary Patterns (LBP) [9] were used with SVM, where LBP was employed to describe faces, then Adaboost was used to select the discriminative LBP features, followed by an SVM for classification. The author reported results on the Labeled Faces in the Wild (LFW) database [10].

Cao et al. [11] presented a shape-based approach where they used topological information extracted from facial landmarks to perform gender classification. The authors compared their technique to Local Binary Patterns which is an appearance-based classifier, and showed a slightly lower performance that is due to the simplicity and small amount of information encoded in the metrology features. In our work we focus on combining shape information with the face appearance for boosted accuracy.

3 Individual Experts and Features

In this section, we shed some light on the classifiers we used as independent experts in our merger, along with the features supplied for each classifier.

3.1 Features

We compared four different types of features; three of which are appearance-based and one is shape-based.

Appearance-Based Approach. We used three types of features that can be extracted from the appearance of a face in the image; which are normalized pixel values (to be in the range $[0 - 1]$), Principal component analysis (PCA) [12], and Haar-like features [13].

Shape-Based Approach. We used the positions of 76 facial landmarks, that were automatically located on the face then their coordinates values were shifted to have the nose-tip at the center. The positions are then normalized by scaling them so that all faces have a constant inter-eyes distance.

3.2 Individual Gender Classification Methods

To perform classification, we chose to use Support Vector Machines (SVM) [7] which are well known for their accuracy and speed. However, for the Haar-like features, due to their high-dimensional vector, we used Adaboost [14] to select the most discriminant features.

For SVM, we specifically use Least-Square SVM (LS-SVM) [15] which are the least squares versions of SVM in which the solution is found by solving a set of linear equations instead of the convex quadratic programming problem for classical SVMs. We use SVM with Radial Basis Function (RBF) Kernel.

4 Proposed Mixture of Experts

We present here the details of our approach for merging several individual classification methods in order to achieve higher gender recognition accuracy.

4.1 Mixture of Experts

A Mixture of Experts (MoE) is a form of dynamic committee machines, where the outputs of the constituent experts (classifiers) are non-linearly combined by some form of gating system to produce an overall output that is superior to that of any single expert alone. In MoE, the input signal is also directly involved in actuating the integration mechanism as shown in Fig. 1.

4.2 Proposed Basic Mixture of Experts with Bayesian Combiner

The Naive Bayes classifier is based on the Bayes' theorem;

$$p(C|F_1, \dots, F_n) = \frac{p(C)p(F_1, \dots, F_n|C)}{p(F_1, \dots, F_n)}, \quad (1)$$

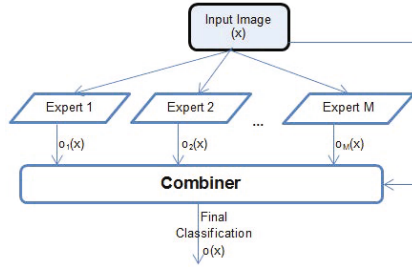


Fig. 1. Structure of a general Mixture of Experts network

where F_1 to F_n are the input features, and C is the class of these features. The denominator of this fraction can be neglected as it does not depend on the class C ; then the theorem can be stated as;

$$posterior \propto (prior \times likelihood) . \quad (2)$$

We adopt this Naive Bayes approach for the merger stage of our mixture machine using the scores as inputs, and C representing the chosen expert. In other words, for each image I and M different experts, the combiner will return the decision of one chosen expert, depending on the set of scores returned by the M experts. The scores are independent of each other given a certain expert, hence the Naive concept.

Training. Each expert is trained using a training subset of the database images. We then train the Bayesian merger using another subset of images. This is done by calculating the prior of each of the contributing individual methods and the likelihood of their outputs' scores as follows:

For each expert, we run its previously-trained classifier on this subset and obtain the following:

- *Prior* This is the classifier's achieved accuracy on this data subset. At merge time we normalize it with the other experts' priors, so that $\sum_{m=1}^M prior(m) = 1$; where $prior(m)$ is the prior of the m 's classifier and M is the number of experts to be merged.

- *Likelihood* We use the scores returned by the classifier from this subset, then split the score range into N intervals. For each interval, we calculate the percentage of images that were correctly classified (true percent) and those that were wrong (false percent). Likelihood of this interval is the true percent minus the false percent, which might result in a negative value if there are more falsely classified samples in an interval than the correct ones. In this case, we shift all the values to have a $min = 1$, and finally, the values are normalized to $[0 - 1]$. For our machine, we choose $N = 20$.

Merging. When a new face is to be classified, each expert, m , is run on the input image returning a binary output o_m along with its score c_m . The prior of each expert alone is retrieved, and these priors are normalized. Then we retrieve the likelihoods of the returned scores, using the trained Bayesian merger. The posterior of each method is then calculated by:

$$posterior(m, c_m) = pr(m) \times likelihood(m, c_m) . \quad (3)$$

The final output (class) is then taken to be that of the method with the highest posterior.

4.3 Proposed 2-Stages MoE

The experts contributing in the MoE should be chosen such that they classify more images differently, which will happen if each expert relies on different cues for its decision.

The effect of increasing the number of experts used in the machine will be discussed in Section 5.4, and it can be expected that using more experts might lead to higher accuracy.

However, we introduce an enhancement to the basic Bayesian MoE proposed in the previous subsection, which allows us to achieve these high correct rates using only two main experts, by adding another stage to this basic MoE, prior to the Bayesian stage.

Using two experts only, will cause the merge machine to be invoked only if each decides the image to belong to a different class.

Training Stage(1). Stage(1) is a classifier trained on the same subset of images used to train the Bayesian MoE (which is now Stage(2)).

Each trained expert is used to classify this subset and return scores for its decisions. The scores for each image are concatenated to its normalized pixel values to form a single vector used to train the classifier of Stage(1). A threshold score $Conf_{Thr}$, is calculated as the average score of the correctly classified faces in this subset.

Merging. If both experts disagree on which class an image belongs to, then their scores are fed to Stage(1) along with the image itself, to return a decision and a score. The decision of Stage(1) supports one of the experts to be the final decision only if its confidence is above the calculated threshold score Sc_{Thr} ; otherwise, its decision is discarded and Stage(2) is used to resolve the conflict based on the experts' posteriors as done in the basic Bayesian MoE. The use of the image itself in the decision of Stage(1) is what separates our MoE from a standard dynamic committee machine.

The structure of the proposed 2-Stages MoE is depicted in Fig 2.

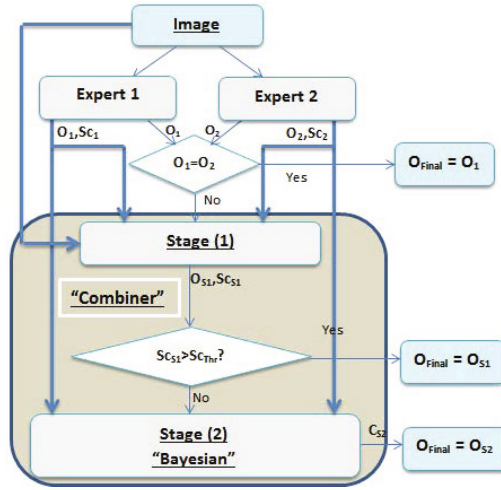


Fig. 2. Structure of the proposed 2-Stages Mixture of Experts network. O is output class with score Sc . Sc_{Thr} is the threshold score of Stage (1).

5 Experiments and Results

5.1 Datasets and Experimental Setup

Our experiments were performed on three face image databases; two of which are well known and publicly available: the FERET image database [5], the LFW database [10]. We have used also a dataset we refer to as MixDB, containing images that were privately collected including several ethnicities; Caucasians, Asians and also some from African descent. We used frontal and near-frontal face images.

While FERET DB contains studio-setting constrained images; LFW offers a unique collection of faces captured from the web, which represents a variation of expressions and lightings. From the FERET database we used one image per subject from the frontal *fa* gallery. For the LFW we selected the frontal faces and formed a set containing 4500 males and 2340 females, having at most two images of the same subject.

For each DB, we created three subsets;

- Training set: used for training individual classifiers;
- Extra-Training set: used for training the combiner stages; and, finally,
- Testing set: used to evaluate the classification performance.

Five-folds cross validation is used for our experiments, where the images are divided into 5 folds, keeping the same ratio between male and female faces; duplicate images of the same subject are placed in the same fold. One fold is used for Testing, two for Training and the remaining two for Extra-Training; this process is repeated five times and the average is reported. The number of faces used for each database is shown in Table 1.

Table 1. Number of faces used for each database (Male/Female)

	FERET	MixDB	LFW
All	600/405	1185/1096	4500/2340
1 Fold	120/81	237/219	900/468

For implementation, images preprocessing, training and testing, we used MATLAB. In all images, the eyes positions were manually located, however in practice, the eyes can be located automatically using active appearance model (AAM). Each image was then rotated so that both eyes lie on a horizontal line, and the face area was extracted to be a square with dimensions relative to the inter-eyes distance. Colored images were transformed to grayscale, then the lighting was enhanced using MATLAB’s built-in function *imadjust* which increases the contrast by remapping the intensity values to fill the entire range of $[0 - 255]$.

5.2 Individual Experts

The experts we use in our mixture machine, adopt two classifiers; SVM and Adaboost, each with different features extracted from the image.

The notations we use, are:

- SVM[Norm]: Normalized image pixels,
- SVM[PCA]: Dimensionally reduced vector using PCA,
- SVM[LM]: Landmarks positions, and
- Ada[Haar]: Haar-features.

For SVM[PCA], we varied the number of principal components (PCs) used from 50 to 300, then tested SVM’s classification on different image sizes, from 16×16 to 40×40 with step 8 pixels per side. We obtained best classification results using 150 PCs regardless of the initial images’ size. So, on the following experiments we will use 150 PCs for SVM[PCA].

For SVM[LM], 76 landmarks’ positions are located automatically using Stasm [16] which is an extended version of Active Shape Model. The coordinates of the landmarks are manipulated as explained in 3.1. Subsequently, SVM is trained on these manipulated coordinates; a vector of size $nLM \times 2$, where nLM is the number of landmarks detected.

We carried out two experiments; the first one regards the choice of the resolution of face images to be used. Then the second experiment tests the performance of our proposed Mixture of Experts.

5.3 Experiment (1): Studying the Effect of Image Resolution on Individual Classifiers

The purpose of this experiment is to investigate the effect of changing the input face image size on the accuracy of the individual classifiers.

Fig. 3 compares the weighted average performance of the experts listed in 5.2, except for SVM[LM] which is independent of the image size. Images are resized

using bi-cubic interpolation. For Haar, due to the very high dimension of the Haar-like feature vector, we stopped at 24×24 images, in which case the Haar-features vector’s dimension is 136,656.

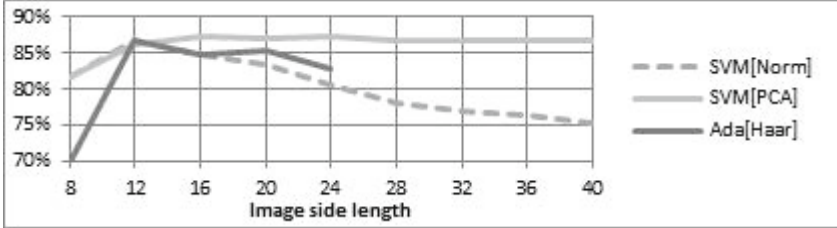


Fig. 3. Classifiers’ responses over different image resolution

From Fig. 3, the following observations are made:

- Increasing the image resolution for the used classifiers does not improve the achieved classification accuracy; but even degrades it significantly when using the image pixels values.

- Using SVM: For a certain higher resolution image; e.g. $40 \times 40 = 1600$, reducing the dimension yields better accuracy; which can be done by two ways; either reducing the image size using simple down-sampling; i.e. SVM[Norm] on lower resolution images, or using PCA; SVM[PCA]. Obviously PCA gave higher performance.

- SVM[PCA]: Its performance is much better than using simple pixel values, and as the number of PCs remain constant (150 in this case), the initial image size does not affect the performance.

- Computational time: Reducing the images’ resolution significantly reduces training time for some experts. For example, on the FERET dataset, Adaboost took 7 minutes for training using Haar features on 400 face images of size 16×16 . When the size increased to 24×24 , the training time jumped to 1 and a half hour.

For the coming experiment we will use low resolution; 16×16 images, since it achieves not only less computational time, but also better accuracy.

Table 2 presents the classification rates of each individual expert on 16×16 images.

5.4 Experiment (2): Proposed MoE

We used our proposed 2-Stages MoE to merge pairs of expert on 16×16 images. Experts contributing in the machine must each be using different features for classification as explained in Section 4, so we specifically chose to merge the shape-based expert, SVM[LM], with each of the other appearance based experts.

For Stage(1), we use SVM as an aiding expert trained with normalized pixel values of the face images concatenated with the score values returned by the two

Table 2. Classification Results of Individual Experts on 16×16 face images, sorted in a Descending Order

Expert	FERET	MixDB	LFW
Ada[Haar]	93.71%	86.89%	88.41%
SVM[Norm]	92.91%	85.58%	88.36%
SVM[PCA]	91.11%	82.68%	87.41%
SVM[LM]	80.19%	77.48%	78.76%

experts. The result of Stage(1) supports one of the main contributing experts' decision only if they both disagree, and the SVM's returned confidence is above the threshold score $S_{C_{Thr}}$.

Table 3. Results of the proposed Bayesian Mixture of Experts on 16×16 images, when one of the experts is shape-based. These results are the weighted average of all databases.

Experts	Best Expert	Basic Bayesian MoE	2-Stages MoE
(1) SVM[LM]+Ada[Haar]	88.59%	90.01%	91.54%
(2) SVM[LM]+SVM[PCA]	86.71%	88.41%	90.51%
(3) SVM[LM]+SVM[Norm]	88.19%	89.66%	-
(4) SVM[LM]+Ada[Haar]+SVM[Norm]	88.59%	90.57%	-
(5) SVM[LM]+Ada[Haar]+SVM[PCA]	88.59%	90.17%	-
(6) SVM[LM]+SVM[PCA]+SVM[Norm]	88.19%	89.74%	-
(7) SVM[LM]+Ada[Haar]+SVM[PCA]+SVM[Norm]	88.59%	90.53%	-

Table 3 presents the best achieved weighted average results using our proposed MoE on 16×16 images, which shows an improvement over the best contributing expert by up to 3%. Best classification rate is achieved using our proposed 2-Stages MoE to merge the shape-based SVM[LM] with the appearance-based Ada[Haar].

By comparing row(1) with row(4), and row(2) with row(6), it is observed that the accuracy of the 2-Stages MoE is about 1% higher than that of the Basic Bayesian MoE, when using 3 experts, two of which are the same experts used in the 2-Stages MoE, and the third expert is SVM[Norm]. Using SVM[Norm] as an expert in this 3-experts Basic MoE is the closest to Stage(1) of the 2-Stages MoE, yet trained on normalized image pixels only (without the confidence values of the other two experts). From this we can say that using the aiding expert (here SVM[Norm]) as a part of the combiner stages in the 2-Stages MoE is better than including it from the beginning as a contributing expert while using Bayesian combiner only for the merge.

Fig. 4 compares the performance of the basic MoE with varying number of contributing experts, with the 2-Stages MoE. From this figure it is seen that increasing the number of experts for the basic MoE beyond three does not improve the accuracy, yet using the proposed 2-Stages MoE does.

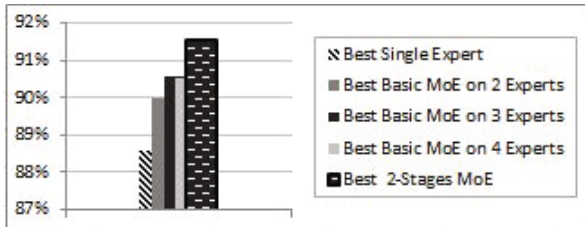




Fig. 4. Best mixture results on 16×16 images. These results are the weighted average of all databases.

Table 4. Best 5-folds Results on each Database

	Best Individual Expert	2-Stages MoE	Other Methods	Classification by humans
FERET	93.71% Ada[Haar]	95.10%±1.2%	- OpenCV: 90.31% - Calfa et al. [6]: 93.95%±2.6%	85.50% 
FERET [4]	91.59%	95.33%	- Mäkinen et al. [4]: 92.86%	
MixDB	86.89% Ada[Haar]	90.12%±3.4%	- OpenCV: 81.67%	77.80%
LFW	88.41% Ada[Haar]	91.49%±1.1%	- OpenCV: 81.58% - Shan [8]: 94.81%±1.1%	86.67% 

5.5 Best Results on each Database

Table 4 presents the best achieved results on each database alone using our proposed 2-Stages MoE for merging SVM[LM] and Ada[Haar]. We compare these results to OpenCV’s gender classification [17] on our subsets, and to published results on the same databases. For FERET, Calfa et al. [6] used the frontal set as we did, and reported results using 5-folds cross validation.

Mäkinen et al. [4] achieved best results using combination by voting of six classification methods, on frontal images with hair. Their subset contained 760 face images divided equally between male and female, from which they used 80% for training and 20% for testing. We report results of our method on their set, splitting the 80% that are used for training equally between the ‘Training’ and ‘Extra-Training’ subsets.

Shan [8] tested SVM on boosted LBP, on the LFW database; a set containing 4500 males and 2943 female, approximately the same number of images we used. They performed 5-folds cross validation for their reported results as we did. We think Shan’s results are better on LFW because of using the boosted LBP which are complicated features compared to the simple features we used for our mixture. However, they have not reported results on any other database.

We also report the average rate of classification done by a small group of 5 people; 3 male and 2 female, on a subset of the 16×16 adjusted face images that was randomly selected from the datasets, 100M/100F from FERET and the same for MixDB, while 150M/150F from LFW. The poor classification rate proves that even though humans can easily classify gender from high resolution face images, it is much harder to classify from very small images like the ones we used.

6 Conclusions and Future Work

In this research, we presented a new approach for merging several gender classification methods using Naive Bayes. We used a Mixture of Experts formed of two stages used to combine results of two experts; the first stage uses the input image along with the scores of the experts for classification, while the second stage implements a Naive Bayes approach for the merge. Our experiments showed that the multilevel MoE composed of only two contributing experts achieves comparable results and sometimes better than using basic MoE which supports more than two experts.

We tested on both appearance-based and shape-based data, and the best merge result obtained are when these were combined together, as shown in Table 3, where the best results are obtained when combining SVM on landmarks positions, with Adaboost on Haar-like features.

The effect of varying the resolution of images on the classification accuracy was studied, from which we proved one of the aspects of the ‘curse of the dimensionality’ problem, showing that higher resolution images are not necessary for better performance. We demonstrated in Section 5.3 that the time taken by the classification process can be reduced a lot without losing much accuracy using low resolution face images.

For future work we plan to use different features extracted from the face image like SIFT, SURF, or boosted LBP along with other classifiers.

References

1. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and Simile Classifiers for Face Verification. In: IEEE International Conference on Computer Vision, ICCV (2009)
2. Ross Beveridge, J., Givens, G.H., Jonathon Phillips, P., Draper, B.A.: Factors that influence algorithm performance in the Face Recognition Grand Challenge. Computer Vision and Image Understanding, 750–762 (2009)

3. Kumar, N., Belhumeur, P.N., Nayar, S.K.: FaceTracer: A Search Engine for Large Collections of Images with Faces. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 340–353. Springer, Heidelberg (2008)
4. Mäkinen, E., Raisamo, R.: An experimental comparison of gender classification methods. *Pattern Recognition Letters* 29(10), 1544–1556 (2008)
5. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The FERET Evaluation Methodology for Face-Recognition Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 1090–1104 (2000)
6. Bekios-Calfa, J., Buenaposada, J.M., Baumela, L.: Revisiting Linear Discriminant Techniques in Gender Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(4), 858–864 (2011)
7. Moghaddam, B., Yang, M.-H.: Learning Gender with Support Faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(5), 707–711 (2002)
8. Shan, C.: Learning local binary patterns for gender classification on real-world face images. *Pattern Recognition Letters* 33, 431–437 (2012)
9. Ojala, T., Pietikainen, M.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(7), 971–987 (2002)
10. Huang, G.B., Ramesh, M., Nick, J., Berg, T., Learned Miller, E.: Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments, Technical report, University of Massachusetts, Amherst (2007)
11. Cao, D., Chen, C., Piccirilli, M., Adjeroh, D., Bourlai, T., Ross, A.: Can Facial Metrology Predict Gender? In: Proc. of International Joint Conference on Biometrics (IJCB), Washington, DC, USA (October 2011)
12. Turk, M., Pentland, A.: Eigenfaces for recognition. *Journal of Cognitive Neuroscience* 3(1), 71–86 (1991)
13. Viola, P., Jones, M.: Rapid Object Detection Using a Boosted Cascade of Simple Features. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 511–518 (2001)
14. Freund, Y., Schapire, R.E.: A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences* 55(1), 119–139 (1997)
15. Suykens, J.A.K., Vandewalle, J.: Least Squares Support Vector Machine Classifiers. *Neural Process. Lett.* 9(3), 293–300 (1999)
16. Milborrow, S., Nicolls, F.: Locating Facial Features with an Extended Active Shape Model. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 504–513. Springer, Heidelberg (2008), <http://www.milbo.users.sonic.net/stasm>
17. OpenCV Gender Classification, <http://docs.opencv.org/modules/contrib/doc/facerec/index.html>