Facial Expression Recognition for Domestic Service Robots

Geovanny Giorgana and Paul G. Ploeger

Bonn-Rhein-Sieg University of Applied Sciences,
Grantham-Allee 20 53757 Sankt Augustin, Germany
{geovanny.giorgana@smail.inf.,paul.ploeger@}h-brs.de

Abstract. We present a system to automatically recognize facial expressions from static images. Our approach consists of extracting particular Gabor features from normalized face images and mapping them into three of the six basic emotions: joy, surprise and sadness, plus neutrality. Selection of the Gabor features is performed via the AdaBoost algorithm. We evaluated two learning machines (AdaBoost and Support Vector Machines), two multi-classification strategies (Error-Correcting Output Codes and One-vs-One) and two face image sizes (48 x 48 and 96 x 96). Images of the Cohn-Kanade AU-Coded Facial Expression Database were used as test bed for our research. Best results (87.14% recognition rate) were obtained using Support Vector Machines in combination with Error-Correcting Output Codes and normalized face images of 96 x 96.

Keywords: Facial expression recognition, Gabor features, AdaBoost, Support Vector Machines, Error-correcting output codes, One-vs-One multi-classification, Face normalization

1 Introduction

Facial expression recognition (FER) offers domestic service robots (DSR) a natural way to interact with humans. This channel of information can be used by robots in order to receive feedback on their executed actions as well as to convey empathy.

In general, changes in a face that are due to the execution of a facial expression can be observed in an image as changes in the face texture. Two-dimensional (2D) Gabor filters have proven success for the representation of the instantaneous appearance of a face via texture analysis. One advantage of such filters is the fact that they can provide a spatially localized frequency analysis of the images. Another strong motivation for its use is its similarity to certain mammal's visual cortical cells [7].

Among the most relevant works studying the performance of Gabor filters for FER we find [2], where authors use local Gabor filter banks and PCA plus LDA for dimensionality reduction. Furthermore, [6] presented a local approach where Gabor features are extracted at the location of eighteen facial fiducial points. A very extensive study is presented in [1], where a comparison of different image

sizes, feature selectors, classifiers and methods to extend them for the multi-class problem is presented.

In this paper, a fully automatic facial expression recognizer that maps the perceived expressions into one of the following three basic emotions: joy, surprise and sadness, plus neutrality is presented. Our approach finds its cornerstone in the normalization of the input images and the extraction of Gabor features. Since the number of dimensions of the initial feature space is very high, AdaBoost has been used for feature selection. Furthermore, we compare and report the performance of the system when normalized images of different sizes, and different binary learning machines in combination with different multi-classification strategies are employed.

The rest of the report is organized as follows: section 2 explains the system and the employed methods, the experiments and the obtained results are reported and analyzed in section 3, finally, section 4 conclude our work.

2 Methods

2.1 System Overview

Training of the whole expression recognition system is accomplished after training N binary classifiers whose individual output will be combined in the testing part by different multi-classification strategies. Figure 1 illustrates the six basic stages of the training process. After the detection of faces and eyes in all images of the training set, the located faces are normalized and a pool of Gabor features is extracted from all of them. The subsequent three steps are performed repeatedly in a series of $t = 1, 2, ..., T$ rounds. For each iteration AdaBoost selects the best feature and the whole system is evaluated against another set of images in order to find out how many features to preserve. The amount of features for which the trained classifier presented better results are preserved.

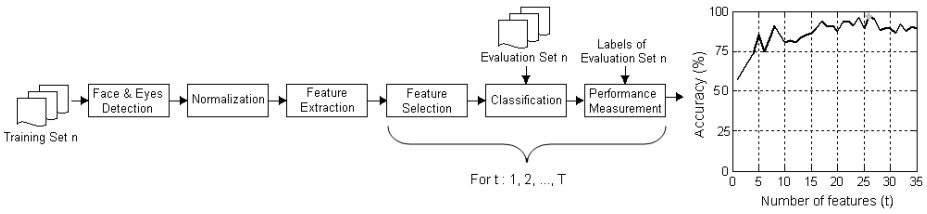


Fig. 1: Diagram of a single binary classifier trainer.

Testing involves similar stages as the training phase (see Fig. 2). After the normalization of the detected faces, only the features selected during training are extracted and used for classification. Classification here is performed by a

set of N binary classifiers whose output are later on integrated by one of the analyzed multi-classification strategies.

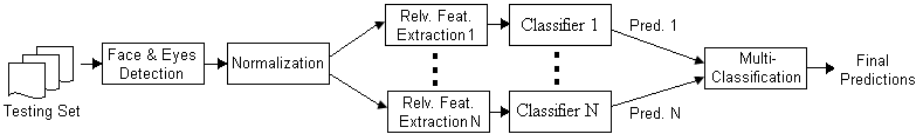


Fig. 2: Diagram of the facial expression recognition system.

2.2 Face Detection and Face Normalization

Face and eyes detection is performed with a commercial software of the company L-1 Identity Solutions, Inc. The system has the feature of fast and robust detection in unconstrained environments. However, the face boundaries are not provided by the software. For this reason all face boundaries were post-generated using a geometric face model that employs the distance between the eyes as primary element [9]. Figure 3a illustrates the characteristics of the model and Fig. 3b shows the face regions obtained after cropping some images of the dataset. As the size of the extracted face regions is not unique, we have scaled all images after cropping. Scaling eases feature-feature comparisons since the center of the two eyes are located at a fixed position, and reduces the number of posterior computations without a drastic loss of accuracy. Finally, the histogram of all scaled images was equalized in order to reduce the effects of illumination. In this work, scaled images of size 48 x 48 and 96 x 96 have been investigated.

(a) (b)

Fig. 3: a) Geometric face model, b) Cropped faces with the model.

2.3 Gabor Feature Extraction

A Gabor filter results from the modulation of a complex sinusoid with a Gaussian envelope. The mathematical form of a complex Gabor function in spatial domain is given in Eq. 1.

$$\psi_{u,v} = \frac{\| \vec{k}_{u,v} \|^2}{\sigma^2} \exp\left(-\frac{\| \vec{k}_{u,v} \|^2 \| \vec{x} \|^2}{2\sigma^2}\right) \left[\exp\left(j \vec{k}_{u,v} \vec{x}\right) - \exp\left(\frac{\sigma^2}{2}\right)\right] \ . \tag{1}$$

In Eq. 1 the parameter $\vec{x} = (x, y)$ represents the position, in spatial domain, of the image pixel where the filter is being applied, σ defines the standard deviation of the Gaussian window in the kernel, $\| \cdot \|$ denotes the norm operator and $\vec{k}_{u,v}$ is the characteristic vector defined as

$$\vec{k}_{u,v} = \begin{pmatrix} k_{(u,v)x} \\ k_{(u,v)y} \end{pmatrix} = \begin{pmatrix} k_v \cos\theta_u \\ k_v \sin\theta_u \end{pmatrix} \ . \tag{2}$$

From Eq. 2, $k_v = \frac{k_{max}}{f^v}$ corresponds to the scale of the filter and $\theta_u = u \cdot \frac{\pi}{8}$ specifies the orientation of the filter.

In principle, the images can be analyzed into a detailed local description by convolving them with a very large number of Gabor filters at different spatial frequencies and orientations. In our work, we created a bank with 40 Gabor filters at 5 different scales and 8 different orientations. Such Gabor filters were tuned with the following parameters: $\sigma = 2\pi, k_{\max} = \frac{\pi}{2}, f = \sqrt{2}, v = \{0, \cdots, 4\}, u = \{0, \cdots, 7\}$.

After creating the bank, the scaled images are convolved with all filters in the bank to obtain their Gabor representation. Eq. 3 shows the mathematical definition of the convolution of an image $I(\vec{x})$ and a Gabor kernel $\psi_{u,v}(\vec{x})$.

$$O_{u,v}(\vec{x}) = I(\vec{x}) * \psi_{u,v}(\vec{x}) \ . \tag{3}$$

The magnitude of the complex function $O_{u,v}(\vec{x})$ is then computed to obtain the Gabor features as follows

$$\|O_{u,v}(\vec{x})\| = \sqrt{\Re^2\left\{O_{u,v}(\vec{x})\right\} + \Im^2\left\{O_{u,v}(\vec{x})\right\}} \ . \tag{4}$$

Finally, all extracted Gabor features from an image are concatenated together to form a feature vector.

2.4 Feature Selection

The dimensionality of the feature vectors is forty times higher than the size of the normalized images. Even small images would bring about very large feature vectors that would either slow down the system or result intractable for most of the existing classification algorithms. Because of that fact, we have employed the AdaBoost algorithm to keep the most useful Gabor features.

AdaBoost is a machine learning algorithm formulated in 1995 by Yoav Freund and Robert Schapire [4]. The learning strategy of this algorithm is based on the Condorcet jury theorem that holds the belief that a group will make better decisions than individuals, given that individuals have a reasonable competence.

AdaBoost aims to build a complex, non-linear "strong" classifier H_T by linearly combining T "weak" classifiers $h_t \in -1, +1$. The algorithm runs in a series of rounds $t = 1, 2, ..., T$, and in each round it chooses the weak classifier that achieved the lowest error in a given training set. A benefical characteristic of the algorithm is the use of two kinds of weights, namely α_t and w_n, where n indicates to what training example w belongs to. On the one hand, the α's weight the h_t's so that the better a weak classifier is the more impact it will have in the last decision; on the other hand, the w's weight all training examples so that in each round the weak classifiers that correctly classified the previously misclassified training examples are favored.

For sake of simplicity, but also due to their proven well performing, decision trees of a single level, also known as decision stumps, have been used as weak classifiers. A decision stump here is defined as a function that evaluates if a Gabor feature $O_m = \|O_{u,v}(\overrightarrow{x})\|$ is above or below a certain threshold λ_m. Eq. 5 shows the mathematical definition of a decision stump, whereas Eq. 6 illustrates how the threshold for each dimension of the feature vector was computed [8]. In Eq. 5, $p_m \in +1, -1$ is a parity to indicate the direction of the inequality and λ_m is the threshold value for dimension m. In Eq. 6, P and Z are the number of positive and negative examples, respectively, whereas $\sum_{p=1}^{P} O_{m|y=1}$ and $\sum_{z=1}^{Z} O_{m|y=-1}$ denote the sum of all Gabor features of the dimension m that belong to positive and negative training examples, respectively. In other words, λ_m corresponds to the mean value of the mean of all Gabor features of dimension m belonging to the positive class and all Gabor features of dimension m belonging to the negative class.

$$h_m = \begin{cases} 1 & \text{if } p_m O_m < p_m \lambda_m \\ -1 & \text{otherwise} \end{cases} \tag{5}$$

$$\lambda_m = \frac{1}{2}\left(\frac{1}{P}\sum_{p=1}^{P} O_{m|y=1} + \frac{1}{Z}\sum_{z=1}^{Z} O_{m|y=-1}\right). \tag{6}$$

2.5 Classification

If AdaBoost were used for classification, the T chosen weak learners together with the T α's would be used as in Eq. 7. However, other classifiers can also be used in combination with the selected Gabor features. We also used Support Vector Machines (SVM) with an RBF kernel to create a hyperplane that optimally separates the different classes of data. In order to tune the parameters of the kernel, we performed Leave-One-Subject-Out Cross Validation (LOSOCV) during training because this process gives us some guarantee for generalization to new subjects and helps us to avoid overfitting.

$$H_T(x_i) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x_i)\right). \tag{7}$$

2.6 Multi-Classification

One-vs-One (1-vs-1) is a strategy to convert binary concept learning algorithms such as AdaBoost and SVM into multi-classification algorithms. For C classes, this strategy creates $k = \sum_{i=1}^{C-1} i$ binary classifiers, each comparing two different classes. When an instance is input to the system for classification, all the 1-vs-1 binary classifiers indicate their belief $y_i \in -1, +1$ and then the class with more votes is declared as the winner.

Error-Correcting Output Codes (ECOC), unlike the 1-vs-1 strategy, offers the possibility to recover from errors made by the individual learners. The first step, as proposed by [3], consists of creating a matrix A whose rows contain unique n-bit code words of zeros and ones. In this case the number of classes C is equal to 4, hence the *exhaustive method* has been used to build A. This method guarantees high inter-row Hamming distance and neither repetition nor complementarity of columns, three necessary conditions to successfully correct errors. The exhaustive method consists of filling row 1 with ones, whereas the remaining rows are filled with alternating runs of 2^{C-i} zeros and ones (here $i = 2, 3, 4$ is the row number). Each word in the matrix represents an expression and has a length $n = 2^{C-1}-1$. The number of errors this scheme can recover from is $\left[\frac{\Delta_{min}(A)-1}{2}\right]$, where $\Delta_{min}(A)$ is the minimum Hamming distance between any pair of codes in A. When an instance is input to the system for classification, the n ECOC binary classifiers indicate their belief $y_i \in 0, +1$ to create a code word that is then compared to the set of words in A according to Hamming distance. Table 1 shows the coding matrix employed for our 4-class problem, whereas Table 2 shows the distribution of the seven created ECOC classifiers.

Table 1: Coding Matrix A.

Expression	**f_0**	**f_1**	**f_2**	**f_3**	**f_4**	**f_5**	**f_6**
Neutral	1	1	1	1	1	1	1
Joy	0	0	0	0	1	1	1
Surprise	0	0	1	1	0	0	1
Sadness	0	1	0	1	0	1	0

Table 2: Distribution of ECOC classifiers.

Classifier	**Classifier distribution**
f_0	neutral vs rest
f_1	neutral-sadness vs rest
f_2	neutral-surprise vs rest
f_3	joy vs rest
f_4	neutral-joy vs rest
f_5	surprise vs rest
f_6	sadness vs rest

3 Experiments and Results

The eight combinations that can be formed with the two normalized face images sizes, the two learning machines and the two multi-classification strategies described in Sect. 2 were tested against images from the Cohn-Kanade AU-Coded Facial Expression Database [5].

In general, the first and the last frame of the sequences of 96 subjects showing expressions of joy, surprise, sadness and neutrality were selected for training and testing. The test set was collected from 46 subjects and was made up of 37 images of each expression. The training set contained images of the remaining 50 subjects; however, the exact number of training images depended on the multi-classification strategy since we took care that each binary classifier were as balanced as possible. The evaluation set contained the same subjects used for training, but the second frame and one frame before the last were used instead. For each case, the maximum number of features selected per classifier was set to 35 in the training phase (i.e., $T = 35$ according to Sect. 2.1).

Figure 4 illustrates the accuracy rate (black segments), the error rate (white segments) and the tie rate (gray segments) accomplished by all evaluated combinations. In the figure we notice the following:

- The highest accuracy rate (81.76%) was achieved by the two combinations using SVM and ECOC.
- The use of SVM resulted in higher accuracy, except for the case when 1-vs-1 and 96 x 96 normalized images were used together.
- Using ECOC always gave better accuracy than using 1-vs-1, when using the same learning machine.
- When using the same learning machine and multi-classification strategy, 48 x 48 provided better or equal accuracy than 96 x 96, unless AdaBoost and ECOC were combined.
- The use of ECOC turned out into a lower percentage of errors and a higher percentage of ties.

The fact that all combinations using ECOC resulted in a lower error rate and a higher tie rate motivated us to solve the cases of uncertainty with the trained 1-vs-1 classifiers. We applied such classifiers only when the combinations using ECOC finished in a tie. We aimed with this to have a tiebreaker that behaves better than random[1].

Figure 5a shows the new accuracy of the combinations using ECOC and 96 x 96 images, while Fig. 5b shows the new accuracy of the combinations using ECOC and 48 x 48 images. Both figures are divided in two parts, each containing two bars. The left part corresponds to combinations using SVM and ECOC, and the right one to combinations using AdaBoost and ECOC. The left bar of each part (black bars) depicts results after breaking ties with 1-vs-1 classifiers that were trained with SVM, and the right bar (gray bars) the results after breaking ties with 1-vs-1 classifiers trained with AdaBoost.

We notice from the figures that:

- All combinations using 96 x 96 images gave better results than those using 48 x 48.

[1] Random choices were not necessary in this experiment, but would have been if the 1-vs-1 classifiers had not been able to break all the ties.

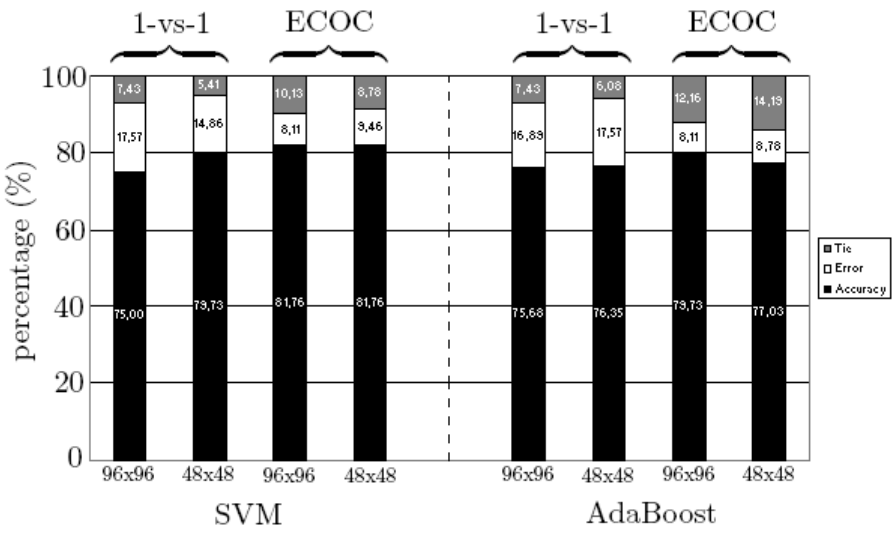
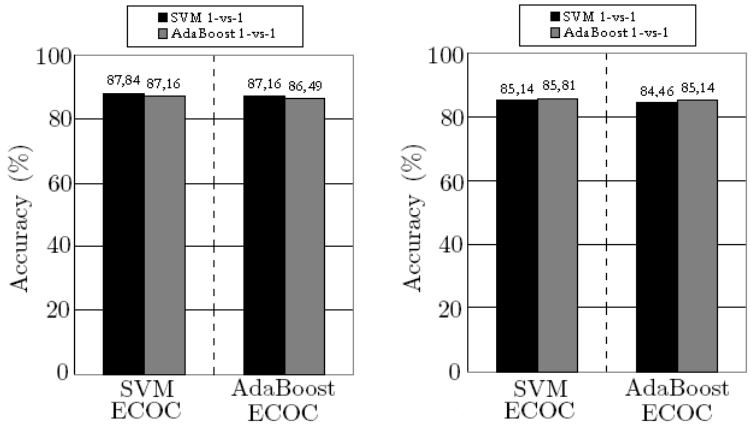


Fig. 4: Accuracy rate, error rate and tie rate of the 8 evaluated combinations.

- SVM ECOC outperformed AdaBoost ECOC for both fashions 96 x 96 and 48 x 48.
- The combination that attained the highest accuracy (87.84%) was the one using the SVM algorithm, the ECOC strategy and normalized images of size 96 x 96, and whose ties were broken with 1-vs-1 SVM classifiers.

Tables 3 and 4 state what is the percentage of ties correctly broken in ECOC systems working with 96 x 96 images and with 48 x 48 images, respectively. The four columns of the tables indicate what is the ECOC system whose ties were broken, the system used to break the ties, what was the total percentage of ties and what percentage of this was correctly classified, respectively. On the one hand, Table 3 reveals that using SVM 1-vs-1 classifiers when the normalized image is of size 96 x 96 results in a higher amount of ties correctly classified than using AdaBoost 1-vs-1 classifiers. On the other hand, Table 4 shows that AdaBoost 1-vs-1 classifiers are more convenient than SVM 1-vs-1 classifiers when the normalized image is of size 48 x 48.

Table 5 presents the confusion matrix of the system with the highest accuracy. We see from the matrix that all neutral expressions were correctly classified, 81% of the faces posing an expression of joy were correctly classified and 19% were incorrectly classified as neutral, 97% of the faces corresponding to surprise were correctly classified and 3% were mislabeled as sadness, 73% of the sad faces were correctly classified, while 27% were classified as neutral. From the table we can see that neutral and surprise are easier to recognize than joy and sadness. Besides, we observe that the majority of the misclassified expressions were confused with a neutral expression.



(a) 96 x 96 normalized images (b) 48 x 48 normalized images

Fig. 5: Accuracy of the combinations using ECOC, after their ties were broken with the 1-vs-1 classifiers.

Table 3: Percentage of ties correctly classified in ECOC systems working with 96 x 96 images.

System to be improved	System used to improve	Ties (%)	Ties correctly broken (%)
SVM ECOC 96 x 96	SVM 1-vs-1 96 x 96	10.13	**60.02**
AdaBoost ECOC 96 x 96		12.16	**61.10**
SVM ECOC 96 x 96	AdaBoost 1-vs-1 96 x 96	10.13	53.30
AdaBoost ECOC 96 x 96		12.16	55.59

Table 4: Percentage of ties correctly classified in ECOC systems working with 48 x 48 images.

System to be improved	System used to improve	Ties (%)	Ties correctly broken (%)
SVM ECOC 48 x 48	SVM 1-vs-1 48 x 48	8.78	38.50
AdaBoost ECOC 48 x 48		14.19	52.36
SVM ECOC 48 x 48	AdaBoost 1-vs-1 48 x 48	8.78	**46.13**
AdaBoost ECOC 48 x 48		14.19	**57.15**

Table 5: Confusion matrix

		Actual			
		Neutral	**Joy**	**Surprise**	**Sadness**
Predicted	**Neutral**	100	19	0	27
	Joy	0	81	0	0
	Surprise	0	0	97	0
	Sadness	0	0	3	73

Figure 6 presents four correctly classified images, whereas Fig. 7 shows three incorrectly classified. These particular misclassified images have in common that they show not very pronounced expressions, which can be the reason for misclassification.

(a) neutral (b) joy (c) surprise (d) sadness

Fig. 6: Four images correctly classified by the system with no ties and highest accuracy.

(a) joy (b) sadness (c) sadness

Fig. 7: Three images incorrectly classified by the system with no ties and highest accuracy. The actual label of (a) is joy, while the actual label of (b) and (c) is sadness.

The features selected for each classifier of the ECOC system are shown in Fig. 8. In the figures, most of the features are nearby the mouth. This latter might be because in the database this is the feature that varies the most from expression to expression. We also see that for some images, the depicted features match reasonable facial regions. For example, image 8c shows the features that separate

surprise from the other expressions; most of the features in this image match the circumference of the open mouth. In image 8d we find features matching the lip corners, which is sensible to separate sadness from the rest.

(a) (b) (c) (d) (e) (f) (g)

Fig. 8: Features chosen by AdaBoost for the 7 ECOC classifiers of the system with no ties and highest accuracy. a) neutral vs rest, b) joy vs rest, c) surprise vs rest, d) sadness vs rest, e) neutral-joy vs rest, f) neutral-surprise vs rest, g) neutral-sadness vs rest.

In total, 170 Gabor features out of 386,640 were used for recognition of the 4 emotions, which represents a 2000-fold reduction (see Table 6).

Table 6: Number of features selected by AdaBoost for each ECOC classifier of the system with no ties and highest accuracy.

Classifiers	**# chosen features**
neutral vs rest	24
joy vs rest	24
surprise vs rest	23
sadness vs rest	27
neutral-joy vs rest	26
neutral-surprise vs rest	30
neutral-sadness vs rest	16

4 Conclusion

This study has focused on exploring the performance of SVM and AdaBoost in combination with ECOC and 1-vs-1 for the problem of facial expression recognition from static images. Two different normalized image sizes have been considered in this analysis, namely, 96 x 96 and 48 x 48. Besides, the use of the 1-vs-1 classifiers to solve the uncertainties generated when the ECOC-based systems cannot find a single winner has been investigated. All evaluated combinations utilized features extracted by convolving Gabor filters at specific positions in the images, and with particular scale and orientation. The experiments revealed

that if the normalized images are of size 96 x 96, the SVM 1-vs-1 classifiers convert more ties into correct classifications than the AdaBoost 1-vs-1 ones. On the contrary, if the normalized images are of size 48 x 48, the AdaBoost 1-vs-1 classifiers are more convenient. Another important result is the fact that after tie breaking, 96 x 96 normalized images turned out in about 2% more accuracy than 48 x 48 ones, and that the use of SVM gave better accuracy than the use of AdaBoost. It is also worth mentioning that before tie breaking, ECOC-based systems resulted in better accuracy. AdaBoost as feature selector allowed us to achieve over 2000-fold reduction. The results also show that all combinations have the characteristic of being person-independent and can perform automatically from end to end with the help of an accurate eye detector as the one provided by the L-1 Identity Solutions, Inc.

References

1. Bartlett, M.S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., Movellan, J.: Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior. In: IEEE International Conference on Computer Vision and Pattern Recognition. pp. 568–573 (2005)
2. Deng, H.B., Jin, L.W., Zhen, L.X., Huang, J.C.: A New Facial Expression Recognition Method Based on Local Gabor Filter Bank and PCA plus LDA. International Journal of Information Technology 11, 86–96 (2005)
3. Dietterich, T.G., Bakiri, G.: Solving Multiclass Learning Problems via Error-Correcting Output Codes. Journal of Artificial Intelligence Research 2, 263–286 (1995)
4. Freund, Y., Schapire, R.E.: A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. In: Proceedings of the Second European Conference on Computational Learning Theory. pp. 23–37 (1995)
5. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive Database for Facial Expression Analysis. In: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG'00). pp. 46–53 (2000)
6. Koutlas, A., Fotiadis, D.I.: An Automatic Region Based Methodology for Facial Expression Recognition. In: IEEE International Conference on Systems, Man and Cybernetics. pp. 662–666 (2008)
7. Kulikowski, J.J., Marčelja, S., Bishop, P.O.: Theory of Spatial Position and Spatial Frequency Relations in the Receptive Fields of Simple Cells in the Visual Cortex. Biological Cybernetics 43, 187–198 (1982)
8. Shen, L., Bai, L.: Adaboost Gabor Feature Selection for Classification. In: Proc. of Image and Vision Computing NewZealand, Akaroa, New Zealand. pp. 77–83 (2004)
9. Shih, F.Y., Chuang, C.F.: Automatic Extraction of Head and Face Boundaries and Facial Features. Information Sciences 158, 117–130 (2004)