Real-Time Gender Based Behavior System for Human-Robot Interaction

Pierluigi Carcagnì[1], Dario Cazzato[1], Marco Del Coco[1], Marco Leo[1], Giovanni Pioggia[2], and Cosimo Distante[1]

[1] National Research Council of Italy, Institute of Optics, Arnesano (LE), Italy
pierluigi.carcagni@ino.it
http://www.ino.it/en/
[2] Institute of Clinical Physiology of CNR, Pervasive Healthcare Center, Messina, Italy

Abstract. This work introduces a real-time system able to lead humanoid robot behavior depending on the gender of the interacting person. It exploits Aldebaran NAO humanoid robot view capabilities by applying a gender prediction algorithm based on the face analysis. The system can also manage multiple persons at the same time, recognizing if the group is composed by men, women or is a mixed one and, in the latter case, to know the exact number of males and females, customizing its response in each case. The system can allow for applications of human-robot interaction requiring an high level of realism, like rehabilitation or artificial intelligence.

Keywords: human-robot interaction, artificial intelligence, gender recognition.

1 Introduction

Each human-human communication is based on a form of interaction that involves faces. In the light of this, for the design of a human-computer interaction, it is natural to expect to find faces playing an essential role. In fact, there has been considerable technical progress within artificial intelligence in the field of computer vision to open the possibility of positioning faces at a very significant place within human-machine interaction [14]. In the field of artificial intelligence and human-robot interaction (HRI), even gender recognition can significantly improve the overall user experience quality, giving to the person the opportunity to interact with an entity that can change its behavior depending on the sex of the user that is interacting with it. Beyond realism and variance of the interaction, a gender recognition system able to work in real-time could lead to several applications in the field of socially assistive robotics, like people in rehabilitation or autistic children, considering their well-known interest on computers and electronic devices [19].

Since its importance, this topic has been well investigated in the last decades by computer vision and machines learning scientists. As a preliminary step, especially in order to create a fully automatic face analysis system, facial images of men and women must be extracted. The well-known Viola-Jones [26] algorithm introduces a robust cascade detector (based on AdaBoost [9] and Haar features) for the face recognition in image, and is actually considered as a state-of-art approach.

M. Beetz et al. (Eds.): ICSR 2014, LNAI 8755, pp. 74–83, 2014.

Gender recognition can be viewed as a two-class classification problem, and methods can be roughly divided in feature-based and appearance-based. Mäkinen and Raisamo [18] and Sakarkaya et al. [22] introduced two wide interesting surveys that exhaustively cover the topic.

The very first results were simultaneously shown in [7] and [10], in 1990. A following study, that investigated the use of geometrical features in order to achieve gender recognition, was performed by Brunelli and Poggio [4](1995), while Abdi at al. [11], in the same year, applied pixel based methods and used a radial-basis function (RBF) network. Lyons at al. used Gabor wavelets with PCA and Linear discriminant analysis (LDA) [17]. In 2002, Sun at al. showed the importance of features selection for generic algorithms [24] first and, successively, tested the efficiency of Local Binary Pattern (LBP) for gender classification [23]. Seetci at al. applied Active Apparence Models (AAM) to this scope [21], with the support of an SVM classifier. Recently, Ihsan et al. showed the performance of a spatial Weber Local Descriptor (SWLD) [25].

The problem of gender estimation, together with all the other information extractable from facial images, as a way to be considered in the design of HRI applications has been taken into account already in [27], but gender has been considered only for the design of humanoid faces, and not as a possibility of improving social interaction thanks to the possibility to perform a recognition task on the user's face. Recently, in [13], performances comparison of gender and age group recognition to carry out robot's application service for HRI has been proposed, but with the usage of audio information only. The work of [16] addresses the same problem, but using a RGB-D device and basing its processing on the body shape.

Although several works on the topic of gender recognition have been proposed over the years, in both academia and industry, it seems that very few applications of it in the field of human-robot interaction have been taken into account. Moreover, the only work of this kind in the state of the art does not explore 2D visual information. To overcome to these limitations, in this work, a real-time system that, processes data coming from a camera on board the robot is automatically able to provide more situation awareness if the person in front of it is a male or a female, is proposed. The system can also manage multiple persons at the same time, recognizing if the group is composed by men, women or is a mixed one and, in the latter case, to know the exact number of males and females, customizing its response in each case. The manuscript is organized as follows: in section 2, our system is presented. After introducing the overall scheme, we will focus on the used gender estimation algorithm. Section 3 shows experimental results. Finally, obtained results and future developments are discussed in section 4.

2 NAO Gender Based Behavior System

In Fig. 1 a scheme of the proposed system is shown. It is composed by two main units: the first unit is the Aldebaran NAO humanoid robot, while the second one is a Remote Computational Unit (RCU) aimed to perform all the computational tasks. RCU and NAO are connected by a local network, as shown in Fig. 1. This architecture allows to satisfy the fundamental requirement to work in real-time, avoiding an overload on the low computational power of the robot CPU (an ATOM Z530).

Video frames, coming from the camera mounted on the top of the head of the robot, are taken by means of the API (Application Programming Interface) provided with the NAO Software Development Kit. Captured video frames are sent to the Gender Prediction Module (GPM) subsystem in order to detect the presence of a human being and predict his/her gender. Gender predictions are then sent to the Behavior Decision Module (BDM) that sends a message to the robot in order to activate gender-specific behaviors.

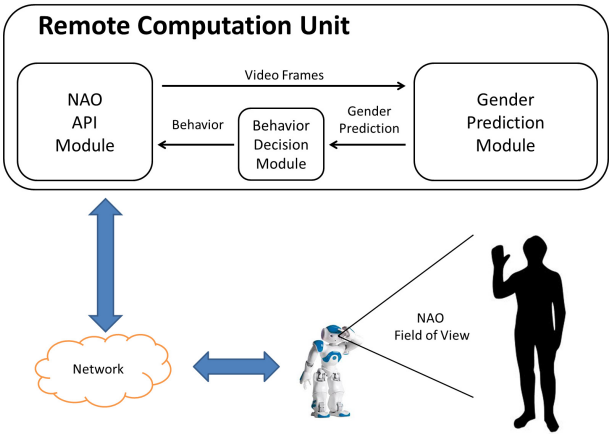


Fig. 1. A scheme of the proposed gender based behavior system

Communication between NAO and the RCU has been achieved using the NAOqi framework, that allows homogeneous communication between different modules (motion, audio, video), homogeneous programming and homogeneous information sharing. After connecting to the robot using an IP address and a port, it is then possible to call all the NAO's API methods as with a local method. For further informations, refer to the official documentation [1].

2.1 Gender Prediction Module

The system core is the *gender prediction module*. It uses the raw video frames as input to detect the presence of a human being and predict his/her gender. As illustrated in Fig.2, the first step is to recognize the presence of a face (consequently a human being) in the scene and to extract the normalized face to analyze. To this end, a *face detection and normalization* process is done by means of the procedure proposed by Castrillon et al. in [5] and the `processed face` is obtained. Moreover, this procedure allows to detect and to track multiple faces in the scene assigning them unique IDs, allowing for particularizing a behavior only one time for a specific person. Once the normalized face image is available, a *features extraction* phase is performed. In particular, we chose to work with Histogram of Oriented Gradients (HOG) that shows, since previous tests, better performance against other low complex features. The procedure aimed to the features extraction is well discussed in section 2.1. HOG `features`

`data vectors` are than projected in a low-dimensional subspace through the *subspace projection* block. Subspace projection makes use of a precomputed *projection model* trained over the features extracted from a dataset of thousands of faces. Successively, the `reduced features data vector` is given to the *SVM prediction* block, that gives as output the gender prediction. As well as the subspace projection, the *SVM prediction* needs a model trained over the reduced features data vector of the same faces dataset. *Subspace projection* and *SVM prediction* blocks are detailed respectively in section 2.1. Predicted genders are stored, frame by frame, in a *predicted gender buffer* of length N_{maj} using a FIFO logic. Finally, the *majority filter* compute the gender class with the greater number of occurrences and give in output the `filtered predicted gender class`.

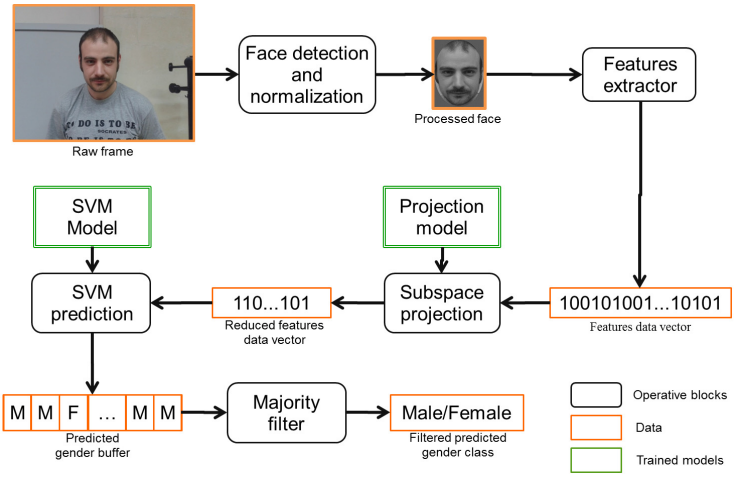


Fig. 2. The block diagram of the gender prediction algorithm: the raw frames are processed in order to obtain a reliable gender-prediction of the people in the scene

Face Detection and Normalization. The detection and normalization of the face in the scene are mainly preprocessing operations whose main steps are illustrated in Fig. 3. It is necessary to guarantee, to the successive operative blocks, a standard face image pose. Castrillon et al. in their face detection and normalization processes [5] perform the sequence of these two operation exploiting persistence face information among successive frames. The current frame is gray-scale converted and then the well known Viola-Jones face detector is applied. Successively, an eye detection is done to locate the eye pairs in the image and rotate and scale the face with the aim to obtain standard face image with eyes pair located in the same position. Down-line the process the result is a normalized 65×59 pixel gray-scale face image.

HOG - Features Extraction. HOG is a well known feature descriptor based on the accumulation of gradient directions over the pixel of a small spatial region referred as a “cell”, and in the consequent construction of a 1D histogram. Even thought HOG

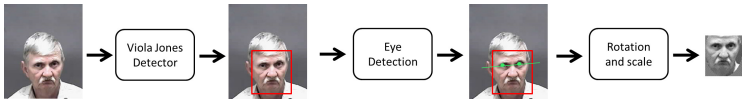


Fig. 3. The face detection and normalization step: the face is cropped and aligned in order to guarantee a standard pose to the *features extraction* step

has many precursors, it has been used in its mature form in Scale Invariant Features Transformation [15] and widely analyzed in human detection by Dalal and Triggs [8]. This method is based on evaluating well-normalized local histograms of image gradient orientations in a dense grid. Let L be the image to analyze. The image is divided in cells (Fig. 4 (a)) of size $N \times N$ pixels and the orientation θ of each pixel $x = (x_x, x_y)$ is computed (Fig. 4 (b)) by means of the following rule:

$$\theta(x) = \tan^{-1} \frac{L(x_x, x_y + 1) - L(x_x, x_y - 1)}{L(x_x + 1, x_y) - L(x_x - 1, x_y)} \tag{1}$$

The orientations are accumulated in an histogram of a predetermined number of bins (Fig. 4 (c-d)). Finally histograms of each cells are concatenated in a single spatial HOG histogram (Fig. 4 (e)). In order to achieve a better invariance to disturbs, it is also useful to contrast-normalize the local responses before using them. This can be done by accumulating a measure of local histogram energy over larger spatial regions, named blocks, and using the results to normalize all of the cells in the block. The normalized descriptor blocks will represent the HOG descriptors.

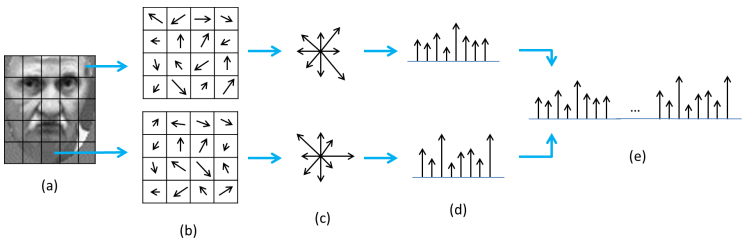


Fig. 4. HOG features extraction: the image is spatially divided in cells and the pixel orientation of each pixel in a cell is computed. Successively orientations histograms are computed and concatenated depending on the cell-space image division.

Subspace Projection. The number of used features for face description is highly influenced in computational complexity and accuracy of classification. Indeed, a reduced number of features allows SVM to use easier functions and to perform better division of clusters. Anyway, the reduction of original features space is a non trivial step.

Principal component analysis (PCA) is a widely used approach for subspace reduction. It chooses a dimensionality reducing linear projection that maximizes the scatter of all projected samples. Simply speaking, the more informative subspace direction are selected for the subspace reduction. The number of components should be selected as the one able to preserve the desired total variance of data.

On the other hand, Linear Discriminant Analysis (LDA) [3] is a class specific method that tries to shape the scatter in order to make it more reliable for classification. This method selects the projection matrix in such a way that the ratio of the between-class scatter and the within-class scatter is maximized. Moreover, in LDA analysis the number of non-zero generalized eigenvalue, and so the upper-bound in eigenvectors numbers, is $c - 1$, where c represents the number of class.

SVM Prediction. Support Vector Machines (SVM)s are techniques aimed to data classification. A classification task uses a training set to generate the model used for the prediction. The training set is usually made up by many instances each of which contains a *class label* and several *features*. The prediction step uses just the features set and the trained model to predict a class for the current instances. As well as for the subspace, projection either the *SVM* accuracy need to be tested over a set of instances different by the training one. At this purpose, in Section 3, a k-fold validation approach has been applied over data.

3 Experimental Results

The *gender prediction module* accuracy evaluation has been realized with a k-fold test over the whole model estimation and prediction process.

We employed a fusion of two of the most representative datasets in face classification problems (on the following referred as "Fusion"): the Morph [2] and the Feret [20] datasets. Both datasets consist of face images of people of different gender, ethnicity and age and are equipped with a complete CVS file with gender, race and other information. Anyway, due to face recognition errors the real number of tested faces is of 55915 male and 9246 females. Even a balanced subset has been taken into account for evaluation. The procedure, showed in Fig.5, consists of two step: a model estimation and a prediction estimation. The whole face-images dataset is randomly split in k subfolds. For each of the k validation steps, $k - 1$ sub-fold for the training and 1 sub-fold for the prediction/validation process have been used. We performed face detection and normalization over each image and successively, the `features data vector` is extracted. The set of `features data vector` is then used to train, in sequence, the subspace reduction algorithm and the SVM prediction one. When both models are available, the one-out fold is tested over them.

The process is repeated over each of 5 to one-out sub-fold combination and the accuracy results is averaged.

For HOG operator, the *VLFeat library*[1] has been used using standard parameters as in [8] with a feature vector length of 2016 elements. Both PCA and LDA subspace projection reduce the features vector dimension. In our case, a number of 100 component for the PCA was taken into account in order to preserve the 95% of the total variance of data. On the other hand, the LDA approach is characterized by a dimension of the projection space that is fixed to 1 (i.e. the number of classes minus one).

The SVM classification problem has been treated by means of the publicly available LIBSVM library [6]. More precisely, we used a radial basis function (RBF) that, in the

[1] `http:www.vlfeat.org`

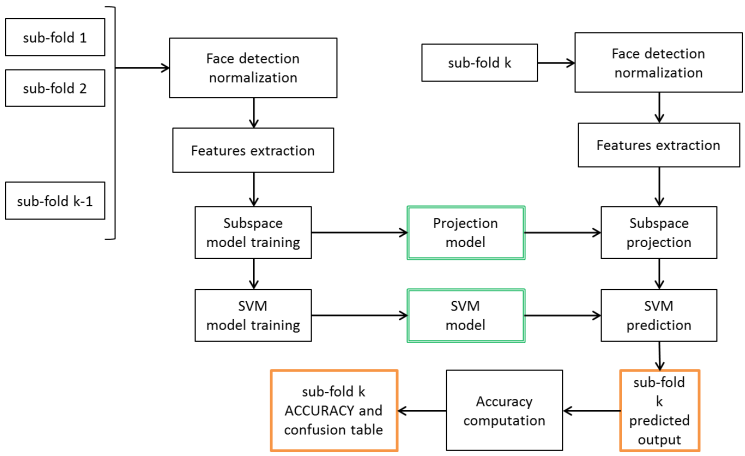


Fig. 5. Test procedure for accuracy estimation: the procedure is done k times in order to obtain the best estimation of total accuracy and confusion table

opinion of the authors of as well as in our experience, seems to be the most reasonable choice [12]. Usually a grid search for penalty parameters C and the others RBF parameters could be desirable. Anyway, our tests does not arise any significant difference in the results as the parameters change. More specifically, we set $C = 1$ and $\gamma = 1/N_f$ where N_f is the number of features.

We obtained a total accuracy of 86.5% and 88.6% for PCA and LDA respectively for the unbalanced dataset, while balanced dataset showed an accuracy of 89.7% with PCA and 80.5% with LDA. Confusion tables are presented in Tables 1 and 2, where M_T, F_T, M_P and F_P represent respectively the true male and female subjects and the predicted ones, TA is the total accuracy and the superscript B stands for *balanced*. All the results are quite close, anyway the HOG+PCA on the balanced dataset gives the best performances both in terms of total accuracy and gap among the two genders.

The whole architecture (presented in section 2) has been tested on a real scenario where people directly interacted with the robot. No constraint in the appearance nor in the background were given to the participants. Each person, one at time, entered in the field of view of the NAO robot. When the face was detected, depending on the gender of the person, the robot acted in a different way. For our purpose, i.e. in order to show the possibility to develop a complete different behavior depending on the user (even originating different learning scheme, since it would be based on the same input), the robot acted in the following way: in the presence of a woman in the scene, it bowed down, while in the presence of a man, the robot greets with his right hand. Fig. 6 illustrate the NAO point of view and the recognize step (a,b) and the consequent action depending on the male (c) or female (d) interacting subject. Even a sentence to be pronounced from the robot has been customized depending on the sex. In the presence of a mixed group (without overlapping of the face area), the robot can say the exact number of men and woman in the scene. Errors are completely related to the errors in the gender prediction algorithm. Moreover, since given a person each prediction is independent from the

Gender confusion tables: each table presents the results for the each specific descriptor/projection pair for both balanced and unbalanced data-set configuration. M_T: Male true; F_T: Female true; M_P: Male prediction; F_P: Female prediction; M_P^B: Male prediction using balanced data-set; F_P^B: Female prediction using balanced data-set; TA: Total accuracy; TA^B: Total accuracy using balanced data-set.

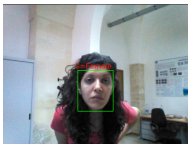
Table 1. HOG + PCA

	M_T	F_T
M_P	97.8%	2.2%
F_P	25.8%	74.2%
M_P^B	87.3%	12.7%
F_P^B	7.8%	92.2%
TA	86%	
TA^B	89.7%	

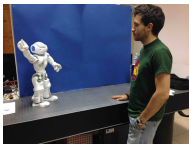
Table 2. HOG + LDA

	M_T	F_T
M_P	98.7%	1.3%
F_P	21.5%	78.5%
M_P^B	82.1%	17.9%
F_P^B	21.2%	78.8%
TA	88.6%	
TA^B	80.5%	

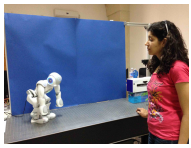
(a) NAO gender recognition step (male).



(b) NAO gender recognition step (female).



(c) Male behavior after recognition.



(d) Female behavior after recognition.

Fig. 6. A test of the interaction between the NAO and humans being. The NAO recognizes the gender of the interacting subject (a,b) and reacts with a customized behavior (it bows down for woman and greets with its right hand for male).

possible presence of other faces in the same image, it was possible to estimate the error of the system evaluating the interaction with the robot of one person at time. With our real scenario, we tested the algorithm on 20 persons, 10 males and 10 females, and 3 errors have been reported. Therefore, the estimated error was of 15%. The system was able to detect and classify faces at a distance in the range of $[20, 300]$ cm.

About computational remarks, the system was tested on a local network in order to avoid latency errors in the evaluation of the frame rate. The RCU was a CPU i7@3.20GHz with a RAM of 16 GB DDR3. Images were processed as a resolution of 640×480. In these conditions, our system was able to work at a frame rate of 13 fps. This is a very encouraging result since it allowed to use the predicted gender buffer in order to strengthen the prediction.

4 Conclusions

With this work, a real-time system able to process data coming from a camera installed into an Aldebaran NAO humanoid robot in order to define, depending on the gender of

the person, its behavior, has been proposed. Multiple persons in the scene at the same time are also managed. The system can allow for applications of human-robot interaction requiring an high level of realism, like rehabilitation or artificial intelligence. A simple customized behavior has been implemented in order to show the possibility to use the system as a starting point for developing a more complex artificial intelligence for the robot, with a more advanced behavior and different tasks. Moreover, other information can be integrated, like an estimation of race and/or age of the users, augmenting the level of the interaction. Additionally, in the case of false prediction, it could be possible to integrate a technique based on gesture recognition in order to, with a pre-specified gesture, teach the robot the right gender of the user, that will store the information. Finally, a user study to investigate whether and how gender-based interaction scheme can improve HRI could be conducted. An evaluation of these developments will be the subject of future works.

References

1. https://community.aldebaran-robotics.com/doc/1-14/index.html
2. Morph-noncommercial face dataset, http://www.faceaginggroup.com/morph/
3. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7), 711–720 (1997)
4. Brunelli, R., Poggio, T.: Hyberbf networks for gender classification (1995)
5. Castrillón, M., Déniz, O., Guerra, C., Hernández, M.: Encara2: Real-time detection of multiple faces at different resolutions in video streams. Journal of Visual Communication and Image Representation 18(2), 130–140 (2007)
6. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011), software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
7. Cottrell, G.W., Metcalfe, J.: Empath: Face, emotion, and gender recognition using holons. In: Advances in Neural Information Processing Systems, pp. 564–571 (1990)
8. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893 (2005)
9. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55(1), 119–139 (1997)
10. Golomb, B.A., Lawrence, D.T., Sejnowski, T.J.: Sexnet: A neural network identifies sex from human faces. In: NIPS, pp. 572–579 (1990)
11. Abdi, H., Valentin, D., Edelman, B., O'Toole, A.J.: More about the difference between men and women: evidence from linear neural networks and the principal-component approach. Neural Comput. 7(6), 1160–1164 (1995)
12. Hsu, C.W., Chang, C.C., Lin, C.J., et al.: A practical guide to support vector classification (2003)
13. Lee, M.W., Kwak, K.C.: Performance comparison of gender and age group recognition for human-robot interaction. International Journal of Advanced Computer Science & Applications 3(12) (2012)
14. Lisetti, C.L., Schiano, D.J.: Automatic facial expression interpretation: Where human-computer interaction, artificial intelligence and cognitive science intersect. Pragmatics & Cognition 8(1), 185–235 (2000)

15. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60(2), 91–110 (2004)
16. Luo, R.C., Wu, X.: Real-time gender recognition based on 3d human body shape for human-robot interaction. In: Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction, pp. 236–237. ACM (2014)
17. Lyons, M.J., Budynek, J., Plante, A., Akamatsu, S.: Classifying facial attributes using a 2-d gabor wavelet representation and discriminant analysis. In: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 202–207 (2000)
18. Mäkinen, E., Raisamo, R.: An experimental comparison of gender classification methods. Pattern Recognition Letters 29(10), 1544–1556 (2008), http://www.sciencedirect.com/science/article/pii/S0167865508001116
19. Moore, D.: Computers and people with autism. Asperger Syndrome, 20–21 (1998)
20. Phillips, P., Moon, H., Rizvi, S., Rauss, P.: The feret evaluation methodology for face-recognition algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(10), 1090–1104 (2000)
21. Saatci, Y., Town, C.: Cascaded classification of gender and facial expression using active appearance models. In: 7th International Conference on Automatic Face and Gesture Recognition, FGR 2006, pp. 393–398 (April 2006)
22. Sakarkaya, M., Yanbol, F., Kurt, Z.: Comparison of several classification algorithms for gender recognition from face images. In: 2012 IEEE 16th International Conference on Intelligent Engineering Systems (INES), pp. 97–101 (June 2012)
23. Sun, N., Zheng, W., Sun, C., Zou, C.-r., Zhao, L.: Gender classification based on boosting local binary pattern. In: Wang, J., Yi, Z., Żurada, J.M., Lu, B.-L., Yin, H. (eds.) ISNN 2006. LNCS, vol. 3972, pp. 194–201. Springer, Heidelberg (2006)
24. Sun, Z., Bebis, G., Yuan, X., Louis, S.J.: Genetic feature subset selection for gender classification: A comparison study. In: IEEE Workshop on Applications of Computer Vision, pp. 165–170 (2002)
25. Ullah, I., Hussain, M., Muhammad, G., Aboalsamh, H., Bebis, G., Mirza, A.: Gender recognition from face images with local wld descriptor. In: 2012 19th International Conference on Systems, Signals and Image Processing (IWSSIP), pp. 417–420 (April 2012)
26. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, vol. 1, pp. I–511. IEEE (2001)
27. Walker, J.H., Sproull, L., Subramani, R.: Using a human face in an interface. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 85–91. ACM (1994)