CHAPTER X

REAL-TIME RECOGNITION OF THE GENDER OF USERS AROUND A SOCIAL ROBOT: PRELIMINARY RESULTS.

Arnaud RAMEY[1] and Miguel A. SALICHS[2]

[1]Robotics Lab, Universidad Carlos III de Madrid;
arnaud.a.ramey@gmail.com
[2]Robotics Lab, Universidad Carlos III de Madrid;
salichs@ing.uc3m.es

Summary: Human-robot interaction is at the core of social robotics. As such, making the robot aware of the users being aroud it is a key functionality. The problem is twofold: user detection and recognition go hand-by-hand.

In this work, we present the development of a gender recognition functionality for the social robot Mopi. Face detection is made thanks to a Viola Jones detector. Gender recognition can alternatively use three popular techniques: Eigenfaces, Fisherfaces, or LBPH. The training set is built thanks to an innovative technique, using automatic image retrieval techniques. Once the classifier trained, the recognition of the gender of a new user does not require any previous knowledge about her or specific training phase. Extensive experimental results are given for verifying the usefulness of the proposed method.

1. Introduction

Social robotics aims at bringing the advances of robotics into the daily life of human users with no or little technical knowledge and making their life easier.

Face recognition is one of the tasks our brain is able to convey from its earliest age. Experiments in (Turati, Cassia, Simion, & Leo, 2006) show that human babies are able to recognize faces after only a few days of life. However, we still don't know much about how the brain analyzes the structure of human faces in order to make so efficiently their recognition. It is hard to know what features are most useful. Inner features, such as mouth, eyes or nose, or outer features, such as hair layout and head shape, can be used.

Social robotics aims at turning robots into everyday life companions for users with little or no technical knowledge. To achieve such a goal, the robot has an imperative need to grow aware of its environment. Among others, it needs to know which users are in its surroundings. Such a knowledge acquisition can be made thanks to face recognition, in a way similar to the one human do.

However, recognizing the face of a user often requires having seen this user beforehand, which triggers some clumsy and cumbersome registration processes. For many applications, only knowing how many users are around the robot and their gender (male or female) is enough. In this article, we will provide a social robot the ability to detect human faces in its vicinity and predict their gender without additional knowledge about these surrounding users.

The article is structured as follows: in part 2, a comprehensive review of the existing solutions is presented. In part 3, we present a robust face detection based on the Viola and Jones detection framework, with a refinement of the results thanks to the depth information. In part 4, an algorithm for gender recognition is presented, based on state-of-the-art algorithms for face recognition. Then, in part 5, we present the results of a comprehensive series of tests that show the validity of the proposed processing pipeline. Finally, its usefulness and the future directions of this work is discussed in part 6.

2. Related work

As underlined in the introduction, detecting and recognizing users is a challenge at the heart of Human Robot Interaction. By giving the robot awareness of the user moves and intentations, it enables a more natural interaction process For instance, in (Leite, Castellano, Pereira, Martinho, & Paiva, 2012), the authors turned a tabletop robot into a chess player for testing interaction with children. Face recognition is made thanks to the embodied webcam, and the robot personalizes its way of playing according to the behavior of the human player.

In (Tomori, 2012), a simple motorized support for a Kinect device, called KATE (Kinect Active Tracking Equipment) is presented. It has two degrees of freedom, which enables the camera device to look at the object of interest. The latter can either be provided by face detection, or the center of the image. The segmentation of the depth images around this center is made by a modified version of the GrabCut method (Rother, Kolmogorov, & Blake, 2004). However, the KATE platform is not mobile, making the segmentation much easier. The authors in (Salcedo, Pena, Cerqueira, & Lima, 2012) have proposed a vision system for a multi-directionnal robot that enables user and object recognition using Eigenfaces (Turk & Pentland, 1991).

Video surveillance applications have similar needs concerning the detection and the tracking of users in a given spatial domain. The problem is challenging for several factors: changing lighting conditions, noise, partial occlusions between users, just to mention a few. Acquiring a better 3D understanding of the environment can be achieved with for instance a set of calibrated cameras, located at strategic positions in the environment . However, their calibration and synchronization turns out to be a cumbersome process. However, the recent introduction of cheap depth imaging sensors such as the Microsoft Kinect into robotics has resulted in a soaring use of depth map for user detection. In (Clapés, Reyes, & Escalera, 2012), multi-modal cues are gathered to obtain a robust estimate of the position of the users. Face detection with Viola and Jones (more extensively presented in part 3) is mixed with user position estimations, obtained thanks to Random Forests.

Other techniques for gender detection were also proposed. In (Childers & Wu, 1991), the analysis of the spectrum of voice sample between male and female voices enable to detect discriminant features between both genders, for instance in bandwidths and amplitudes.

However, none of these techniques have been integrated into a social robot for real time processing and adapting the behavior of the robot to this information. In the following parts of the article, we will explain how we tackled this challenge

3. Robust 3D face detection

In this part, we will explain how we robustly detect human faces in the video stream of the robot. First, faces are detected in to the color video stream. The depth information is then used for discarding false positives.

Viola Jones detector.

Face detection consists in determining if human faces are visible in a video stream, and if it is the case, what are their position in the image. It is one of the classical challenges in vision. Nowadays, some more or less standard techniques are available and give a very good accuracy rate. The technique presented by Viola and Jones in several articles (Viola & Jones, 2004) is one of the most popular and allows a high framerate. The Viola-Jones object detection framework can actually be trained to detect any object, but it is especially popular for face detection.

The method of Viola and Jones is an example of supervised learning. It first requires a training phase, in which it learns features from a training set of images, some of them containing images, other not. Then, in the detection phase, these features can be applied to determine if a sample image contains faces or not.

The learning is based on the appearance of the training images. The process consists in seizing the content of each image by computing so-called *characteristics* in rectangular zones of the image that overlap each other. These characteristics are a synthetic and descriptive representation of the values of the pixels, and are more efficient to be dealt with. They characterize the difference of sum of pixels of values of adjacent rectangular zones of the image. Some of them are visible in figure 1. In order to be able to compute these characteristics quickly, the authors introduce the concept of *integral image*, which is an image the same dimensions than the original one, where each pixel P contains the sum of the pixels of the original image located up and left of P. The computation of a characteristic over two zones then needs at most six accesses to the integral image values, and hence a constant time.

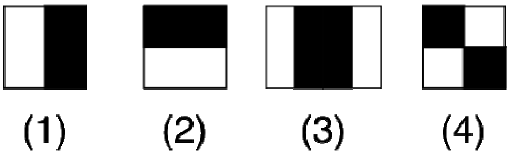


Figure 1: Some of the characteristics used by Viola and Jones .

The second key element in Viola and Jones is the use of a boosting method in order to select the best characteristics. Boosting is a technique that enables the building of a *strong* classifier with a linear combination of *weak* classifiers. In this method, characteristics are seen as weak classifiers. The learning process of the weak classifier hence only consists in learning the threshold value of the characteristic so as to split better positive samples from negatives. The original detector uses three different characteristics, while the modified Lienhart and Maydt (Lienhart & Maydt, 2002) detector adds too other, and includes two diagonal orientations.

For the detection, the classified structure of the boosted classifiers enable a fast detection: Viola and Jones obtain on a Pentium III @ 700 MHz an average processing time of 67 milliseconds, which enable real-time processing on a video feed and is much faster than then-popular similar methods. The strength of the detection is that wide areas of the test image, when marked as negative, are decimated in the first steps, with relatively little data processing.

False positive detection with depth information.

However, the classical Viola Jones detector only uses the color (RGB) image data. As our robot uses a Kinect device, we also have depth data available. that is, zones of the image that are incorrectly classified as faces by the detector.

The underlying idea is the following: the 3D points corresponding to a face obey certain geometric constraints, especially in the width and height of their bounding box. Indeed, two points belonging to one given face cannot be away one of the other of, say, more than one meter.

As such, to determine if a zone of the image classified as a face by the Viola-Jones detector is really a face, we will sample a given number of 2D points from this zone, that we will reproject to 3D using the depth (distance) image. If the bounding box of these reprojected points does not comply with generic given geometric constraints, this detection is classified as a false positive and discarded. The pipeline is illustrated in figure 2. A sample is visible in figure 3.

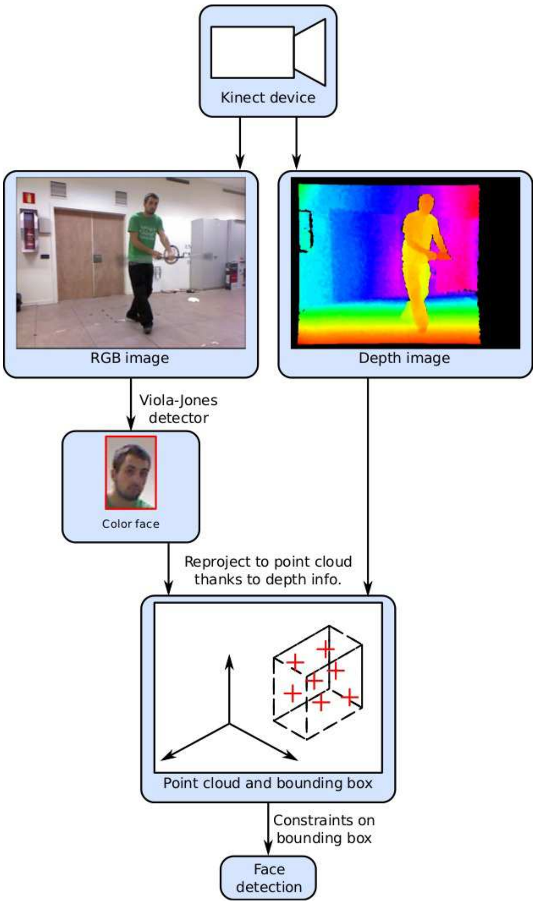

Kinect device
RGB image
Depth image
Viola-Jones detector
Color face
Reproject to point cloud thanks to depth info.
Point cloud and bounding box
Constraints on bounding box
Face detection

Figure 2: Processing pipeline for the face detection algorithm.

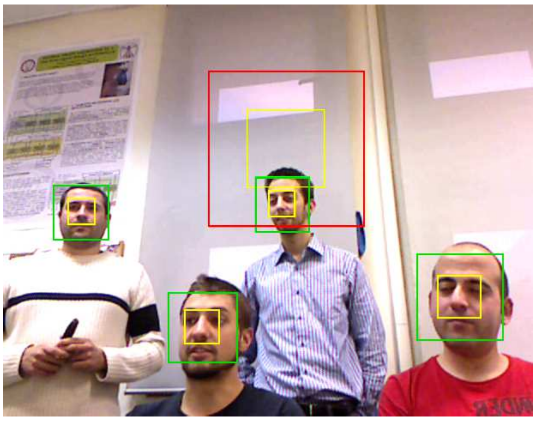


Figure 3: Results of the face detection algorithm on a sample image. The red and green rectangles correspond to the faces as detected by the Viola-Jones detector. The inner yellow boxes are the zones where the sample 2D points are being reprojected in 3D. The 3D-points bounding boxes of the correct detections pass the geometric constraints of a normal 3D face, hence their green color. The 3D-points bounding box of the wall is too big in dimensions and hence is discarded.

4. Gender detection

In previous part 3, we saw how to robustly detect faces in the RGB and depth stream of the robot. We now aim at classify these faces according to their gender (male or female).

Performing gender recognition is a very similar challenge to performing face recognition. The latter consists in determining than a given face image correspond to a given user label. These labels are part of a so-called *training set* consisting of images and the associated labels, for instance 10 pictures of user A, 15 of user B, etc. As such, gender recognition can be seen as a specific case of face recognition with only two meta-users: user *Female* and user *Male*. The training set would then be a collection of pictures of male faces, and another one of female faces.

In section 4.a, we will review the most popular techniques for face recognition. Then, in section 4.b, we will see how a gender faces database can be quickly constituted thanks to automatic retrieval techniques.

a. Face recognition basics

A facial image can be seen as a point in a high-dimensional space. Let us consider images of $p \times q$ pixels. Then each face corresponds to a point in a $p\,q$ -dimensional space. However, for typical dimensions, such as $p=q=100$, this very highly dimensional space makes the matching problem very difficult.

Eigenfaces

The idea behind Eigenfaces is to determine the dimensions that matter the most, that is, that convey the most discriminative information between faces. To achieve this, the theory of Principal Component Analysis (PCA) is used. An extensive review of PCA is available in (Duda, Hart, & Stork, 1995). PCA helps transforming a set of numerous variables possibly correlated into a smaller set of variables that are uncorrelated, and which are the most meaningful for the description of the set. The PCA method finds the directions in the set with the greatest variance in the data, called prin

The Eigenfaces approach is based on PCA.

Here is a brief summary on how to compute Eigenfaces for face recognition:

1. A small set of characteristic pictures is used to train the classifier, which learns how to differentiate pictures thanks to the data distribution of these faces. More exactly, the classifier extracts eigenvalues and eigenvectors from the covariance matrix of the distribution of these training faces. Then, the most discriminative eigenvectors are the ones with the eigenvalues with the highest norm. The number of eigenvectors that should be kept heavily depends from the data, but some rules of thumb are available (Zhao, Chellappa, Phillips, & Rosenfeld, 2003).
2. Faces are then represented as linear combinations of these eigenvectors, called eigenfaces. The eigenface subspace is then defined as the spanned by these eigenfaces. This enables use to make the dimensionality reduction that we sought.
3. Face recognition is then made by projecting a new image into the Eigenface subspace and classifying this new image position in this subspace with relation to the labeled training sample images. The closest neighbor is the most probable recognized face.

Fisherfaces

The idea of PCA, used in Eigenfaces, is simple: find linear combinations of components in the training sample that maximize the total variance in the data. However, it doesn't take into consideration the classes of the training set, and it might happen that found components are not relevant to discriminate between classes of objects. This is the case when most of the variance in the data is not brought by the class itself, but for instance light conditions. In such a case, images projected into the eigenface subspace do not form distinct clusters, and a successful classification becomes challenging.

On the other hand, Fisherfaces also performs a dimensionality reduction, but with respect to the classes of the training samples. As such, the components it computes maximize the inter-class variance, while minimizing the variance within samples of the same class. It learns a class-specific transformation matrix, so that it will find the facial features to discriminate between the classes.

This strategy, called Linear Discriminant Analysis, was presented by the first time by the statistician Sir R. A. Fisher to classify flowers (Fisher, 1936). The first published use of this technique for face recognition was in (Belhumeur, 1997), and the details of the computations of Fisherfaces are available there.

LBPH

Both Eigenfaces and Fisherfaces follow the same idea: a face picture can be seen as a point in a high dimensional space. Similar faces are seen as neighbors in this high dimension space. However, such a high dimension makes it impossible to process in practice. This is why both perform a dimensionality reduction which tries to discriminate between faces, Fisherfaces having the property of making this with respect to the class of the objects. However, both of them remain very sensible to variance sources that do not come from the classes, such as lighting conditions, orientation, pose, etc.

On the other hand, Local Binary Pattern (LBP) is a class of features that is thought to describe the image locally. It finds it roots in texture analysis. The basic idea of LBP is to describe the nature of each pixel (for instance if it is a corner or an edge) with relation to its neighbor pixels. To do that, consider a pixel of the image and threshold its 8 immediate neighbor pixels with its value: if one is higher, denote it with 1, else 0. This gives a pattern of 8 binary values, called *LBP code*, which can be converted into a decimal value. This operator hence transforms an image into a LBP image. By definition, the LBP operator is invariant to linear transforms on the image.

LBP images can then be used for face recognition: the technique presented in (Ahonen, Hadid, & Pietikäinen, 2004) divides an LBP image into n chunks. For each of them, a histogram is computed. The long histogram, obtained by concatenation of these n histograms, constitutes a descriptive feature of the image. Using histogram distances on these long histograms enable a robust and efficient recognition of the class of the image.

b. Building a database of training sample images

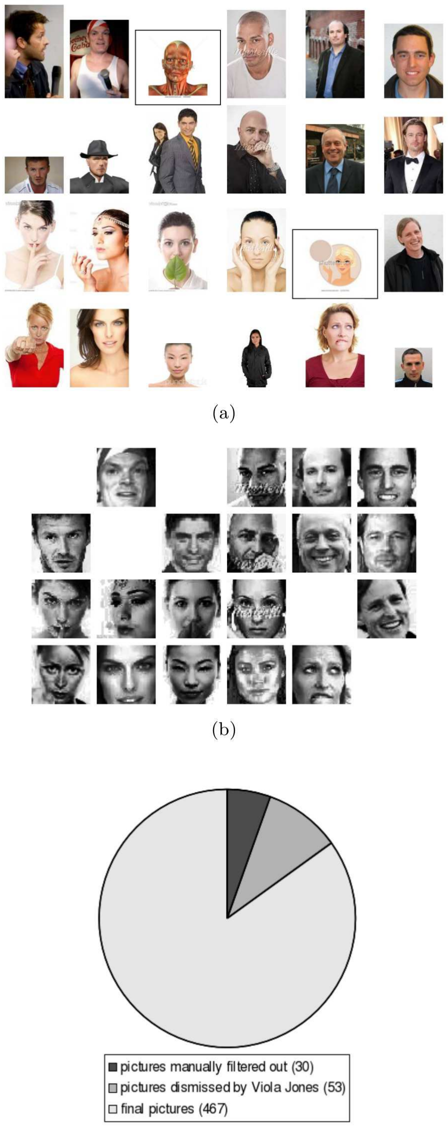
A wide range of face images are available for academic use. For instance, one of the earliest and most used databases is the ORL face database from AT&T Laboratories, Cambridge University, consisting of images of 40 subjects (36 men and 4 women, 10 images per subject) (Samaria & Harter, 1994). Nowadays, an important effort has been made for building extensive face databases, including both an important number of subjects and a variety of pictures per subjects, with varying poses, lighting positions, face expressions, etc. For instance, the Yale database (Lee, Ho, & Kriegman, 2005) contains 5760 single light source images of 10 subjects each seen under 576 viewing conditions (9 poses x 64 illumination conditions).

However, in the scope of this article, we focus only in gender recognition and not in specific face recognition, we decide to constitute a database in an innovative way. A first raw database of 550 images is obtained thanks to an image search engine. We collect the 225 first images returned by Google Images Search with the search query *man face* on the one hand, with the query *woman face* on the other hand.

A first manual review of the so-obtained 550 images allow us to remove the most evident outliers from these raw results, such as pictures not containing faces. The number of images remaining after this quick manual filter is 520.

Then, a second, more accurate filter is made by applying the face detection algorithm described in part 3. The pictures which do not pass this filter might still contain a face, but at least all pictures which pass it do contain an usable picture. Furthermore, the results of the filter supplies the bounding box of the face in the image. As such, it can be used for an exact training of the gender classifier described in 4.a. The number of images remaining after this final filter is 467.

A few samples of the database are visible in figure 4.



(a)

(b)

■ pictures manually filtered out (30)
▨ pictures dismissed by Viola Jones (53)
□ final pictures (467)

(c)

Figure 4: Gender images database constitution thanks to Google Images.
(a): Raw image results from Google Images, with queries man face (two first rows) and woman face (two last rows). The images with a black frame were manually filtered, as they do not contain faces or inappropriate ones.
(b): Remaining images in the database after applying the Viola Jones detector and pre-processing the images. The missing pictures correspond to images that were either discriminated by the manual filtering, or the Viola Jones face detector.
(c): Number of pictures in the database after each step.

5. Implementation for the social robot Mopi

a. Description of the robot Mopi

Hardware specifications

The target robot for this application is the social robot *MOPI* (non-definitive name). It is a home-brew robot of the RoboticsLab of University Carlos III of Madrid shaped as a mobile, car-like platform. It is most notably equipped with a Microsoft Kinect device tiled at 45º, and a (non tiltable) Hokuyo laser scanning range finder. The communication is made through to a WIFI connection.

Software specifications

The robot MOPI works according to the AD paradigm, as presented in (Barber & Salichs, 2002). This paradigm handles skills relying on primitives. Primitives are in direct communication with the physical devices of the robot, and send elementary orders to them. This includes the base motors, the laser sensor, the camera, etc. A skill is the ability of the robot to do a specific action. It relies on the data supplied by the primitives. The actions generated by a skill can be numerous: move the car to a given point, play games with the user, interact with electric appliances, etc.

Since it is an experimental platform, the robot MOPI has been used as a bridge between the traditional implementation of AD, as seen in (Rivas, 2007), and a new one relying on the communication mechanisms of *ROS*, the Robot Operating System (Quigley et al., 2009). The version used for this application is ROS *Electric* running on top of Ubuntu 10.10. The use of ROS most notably gives the possibility to redistribute the computational workload. We can send easily via the wire or a wireless connection raw data from the sensors to remote computers for processing. The latter can be more powerful than the robot embodied PCs, also lightening the computation workload of the main computer. Then, the processed data is sent back to the robot. It also embeds the Stage simulator (Gerkey, Vaughan, & Howard, 2003), which gives us the possibility of making first outlines of our algorithms, for instance the tracking one, before trying them on the real robot.

b. Implementation of gender recognition into Mopi

Both face detection seen in section 3 and robust gender face recognition seen in section 4 have been implemented into the social robot Mopi.

They are structured as two different AD skills running at the same time as background skills into the software architecture of Mopi. The face detection skill subscribes to the image and depth stream coming from the Kinect, and publishes the results of the face detection, containing the rectangular cutouts of the image corresponding to the faces.

The gender recognition skill subscribes to this information and performs gender recognition on the cutouts. The resulting estimated number of male and female users around the robot is then published, and any other skill of the robot can then subscribe to it. A sample is visible in figure 5.

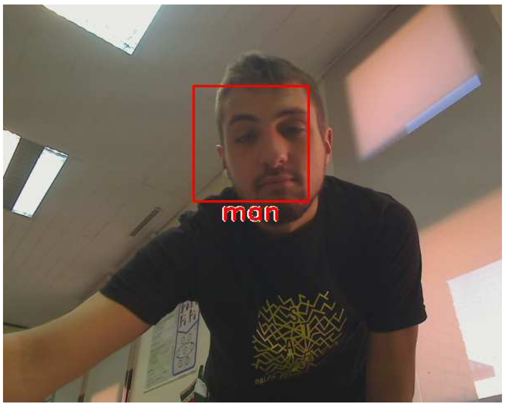


Figure 5: A sample image of the gender recognition skill.

The required times for data processing are showed in figure 6. A typical VGA image, of size $640\times480=3k$ pixels, requires about 70 milliseconds to be procesed. The most costful step is the face detection, the gender classifier running in less than one millisecond.

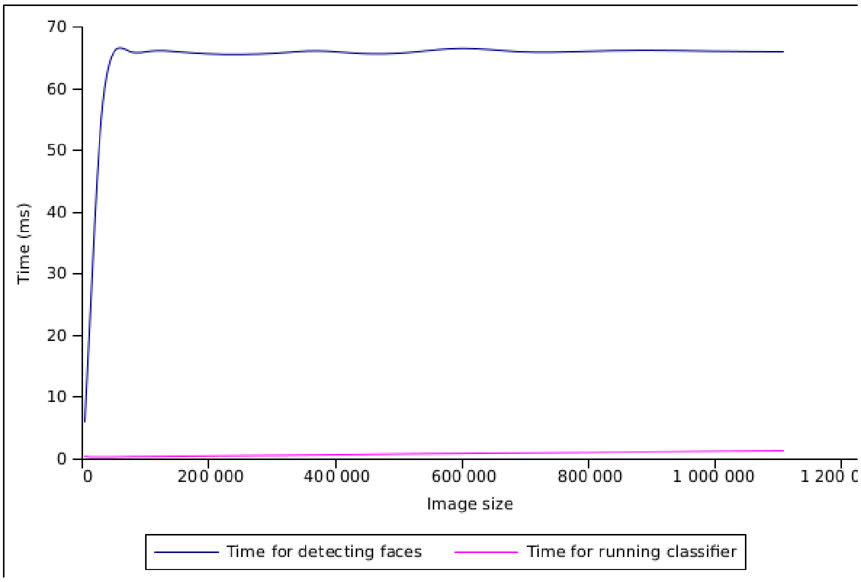


Figure 6: Times needed for detecting faces on a sample image according to the size of the image (number of pixels).

6. Conclusions and future works

In this article, we have presented a method for robustly detecting faces and identifying their gender in the data stream supplied by a range imaging device such as Microsoft Kinect. The face detection is made thanks to Viola and Jones object detecting framework. False positive detections are discarded using depth data and geometric constraints on the points of the possible face once reprojected to 3D. Gender recognition is based on face recognition techniques, using a new face image database made thanks to automatic retrieval techniques. The technique resulting in the most accurate gender recognition has proven to be Fisherfaces.

This method has been succesfully implemented in the social robot Mopi. It enables the detection of the users around Mopi as a skill. The information is then shared to the rest of the software architecture using its inner communication paradigms. The framework is light enough to have it running in the background and supplying it whenever needed by other skills.

Future works will aim at making the gender recognition techniques more robust. Even though Fisherfaces provide a reasonable error rate, better results could be achieved by amplifying the training set. This can be made by adding more pictures for instance from the search engine previously used, or by generating new ones from the existing pictures, for instance by changing their contrast and brightness. Furthermore, the development of higher-level skills that make use of this gender information is ongoing.

7. Acknowledgement

The authors gratefully acknowledge the funds provided by the Spanish MICINN (Ministry of Science and Innovation) through the project *"Applications of Social Robots / Aplicaciones de los Robots Sociales"*.

8. References

Ahonen, T., Hadid, A., & Pietikäinen, M. (2004). Face recognition with local binary patterns. *Computer Vision-ECCV 2004*. Retrieved from http://www.springerlink.com/index/P5D9XP9GFKEX5GK9.pdf

Barber, R., & Salichs, M. (2002). A new human based architecture for intelligent autonomous robots. *Proceedings of The 4th IFAC Symposium on Intelligent Autonomous Vehicles* (pp. 85–90). Elsevier. Retrieved from http://scholar.google.es/scholar?cluster=10839062160608396845&hl=es&as_sdt=2000#0

Belhumeur, P. (1997). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=598228

Childers, D. G., & Wu, K. (1991). Gender recognition from speech. Part II: Fine analysis. *The Journal of the Acoustical Society of America*, *90*(4), 1841. doi:10.1121/1.401664

Clapés, A., Reyes, M., & Escalera, S. (2012). Multi-modal User Identification and Object Recognition Surveillance System. *Pattern Recognition Letters*. Retrieved from http://www.sciencedirect.com/science/article/pii/S0167865512004047

Duda, R., Hart, P., & Stork, D. (1995). Pattern Classification and Scene Analysis 2nd ed. Retrieved from http://www.svms.org/classification/DuHS95.pdf

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Human Genetics*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1469-1809.1936.tb02137.x/abstract

Gerkey, B., Vaughan, R. T., & Howard, A. (2003). The player/stage project: Tools for multi-robot and distributed sensor systems. *Proceedings of the 11th international conference on advanced robotics* (pp. 317–323). Portugal. Retrieved from http://robotics.usc.edu/~gerkey/research/final_papers/icar03-player.pdf

Lee, K., Ho, J., & Kriegman, D. (2005). Acquiring linear subspaces for face recognition under variable lighting. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1407873

Leite, I., Castellano, G., Pereira, A., Martinho, C., & Paiva, A. (2012). Modelling empathic behaviour in a robotic game companion for children. *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction - HRI ’12* (p. 367). New York, New York, USA: ACM Press. doi:10.1145/2157689.2157811

Lienhart, R., & Maydt, J. (2002). An extended set of haar-like features for rapid object detection. *Image Processing. 2002. Proceedings. 2002 International Conference on*. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1038171

Quigley, M., Gerkey, B., Conley, K., Faust, J., Foote, T., Leibs, J., Berger, E., et al. (2009). ROS: an open-source Robot Operating System. *ICRA Workshop on Open Source Software*. Retrieved from http://pub1.willowgarage.com/~konolige/cs225B/docs/quigley-icra2009-ros.pdf

Rivas, R. (2007). Robot skill abstraction for ad architecture. *6th IFAC Symposium on Intelligent Autonomous Vehicles*, *47*(4), 12–13. Retrieved from http://roboticslab.uc3m.es/publications/iav07_AD.pdf

Rother, C., Kolmogorov, V., & Blake, A. (2004). Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (TOG)*. Retrieved from http://dl.acm.org/citation.cfm?id=1015720

Salcedo, C. M., Pena, C. A., Cerqueira, J. de J., & Lima, A. M. N. (2012). Designing a Real Time Artificial Vision System for Human Interaction with an Omnidirectional Mobile Platform. *2012 Brazilian Robotics Symposium and Latin American Robotics Symposium* (pp. 21–26). IEEE. doi:10.1109/SBR-LARS.2012.11

Samaria, F., & Harter, A. (1994). Parameterisation of a stochastic model for human face identification. *Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on*. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=341300

Tomori, Z. (2012). Active Segmentation in 3D using Kinect Sensor. *20th WSCG International Conference on Computer Graphics, Visualization and Computer Vision*. Retrieved from http://wscg.zcu.cz/wscg2012/cd-rom/short/C59-full.pdf

Turati, C., Cassia, V. M., Simion, F., & Leo, I. (2006). Newborns’ face recognition: role of inner and outer facial features. *Child Development*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1467-8624.2006.00871.x/full

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of cognitive neuroscience*. Retrieved from http://www.mitpressjournals.org/doi/abs/10.1162/jocn.1991.3.1.71

Viola, P., & Jones, M. (2004). Robust real-time face detection. *International journal of computer vision*. Retrieved from http://www.springerlink.com/index/q70v4h6715v5p152.pdf

Zhao, W., Chellappa, R., Phillips, P. J., & Rosenfeld, A. (2003). Face recognition: A literature survey. *ACM Computing Surveys*, *35*(4), 399–458. doi:10.1145/954339.954342