Real–Time Face Recognition for Human–Robot Interaction

Claudia Cruz, L. Enrique Sucar and Eduardo F. Morales
Instituto Nacional de Astrofísica, Óptica y Electrónica
Luis Enrique Erro #1, Tonantzintla, Puebla, Mexico
{ccruz,esucar,emorales}@inaoep.mx

Abstract

The ability to recognize people is a key element for improving human-robot interaction in service robots. There are many approaches for face recognition; however, these assume unrealistic conditions for a service robot, like having an image with a centered face under controlled illumination. We have developed a novel face recognition system so that a mobile robot can learn new faces and recognize them in real–time in realistic indoor environments. It is able to learn on-line a new face based on a single frame, which is later used to recognize the person even under different environmental conditions. We employ a preprocessing step to reduce the effect of different illumination conditions, and then identify 3 regions in the face: left eye, right eye and nose–mouth. SIFT features are extracted from each region and stored in a feature vector, which is used for recognition. The matching strategy is able to discard unknown faces and the recognition process uses a Bayesian approach over several frames to improve accuracy. Experimental results in natural environment and with Yale's face database show that that our approach is able to learn different faces from a single image and recognize them on average in three seconds, with very competitive results.

1. Introduction

The ability to recognize and remember individuals is a fundamental requirement for complex interactions among people and represents a key element to improve human-robot interaction in service robots. Several approaches have been proposed for face recognition with various degrees of success (*e.g.* [5,12]); however, most of them assume unrealistic conditions for a robotic framework, like having an image with a centered face with controlled illumination. Our research is oriented towards providing a robot with the capability of learning new faces on-line and recognizing them in real-time in unstructured environments. It is also desirable to reduce false recognition over unknown faces.

We use a face detection algorithm [13] and a simple technique to track the faces in a video sequence. We employ a preprocessing step to reduce the effect of different illumination conditions. We propose a novel face recognition algorithm based on SIFT features [6] over three distinctive regions in the face: left eye, right eye and nose–mouth, and stored them in a vector with a recognition threshold value for each face. The matching strategy is able to discard unknown faces, and the recognition process uses a Bayesian approach over several frames to improve accuracy. Our approach is able to learn on-line a new face based on a single frame, which is later used to recognize the person even under a different viewpoint and environmental conditions. Experimental results with 10 persons in an office environment show that that our approach is able to learn different faces and recognize them on average in three seconds, with very competitive results over existing approaches with more restricted conditions. We also performed experiments on Yale's face database [4] with up to 28 different persons under different illumination conditions.

2. Related work

Face recognition has been an important area of research in the last 30 years. We briefly review the main approaches and then the recent work more related to ours.

The most widely used approaches for face recognition are methods based on dimensionality reduction, such as [12]. These methods consider the face as a whole and try to localize the core components of the face that makes it distinct. Under controlled conditions they provide good results, but often are sensible to rotation and scaling, and require a structured environment. Other techniques use local features for face recognition. An example is [5], which is based on the Gabor wavelet transform. It computes Gabor filtered images in different scales and orientations, which makes the face recognition system more robust under different conditions; however the computational requirements are too high for real–time recognition. Other systems based on the SIFT transform are described later.

978-1-4244-2154-1/08/$25.00

Most face recognition systems are focused on biometric applications, and consider an image of a face under controlled conditions (illumination and view point) which has been previously segmented; so they are not adequate for human–robot interaction. Also, in these systems, the time response is not considered critical. There are some approaches that consider a real–time response [10, 11] or recognition under different conditions [2, 9].

A recent proposal for recognizing persons in videos includes both aspects [1]. They divide the face recognition problem into three main stages: detection, tracking and identification. In the first stage they use a cascade of face detectors [13]. A set of images of the new person are collected along the following frames. When a number of samples have been reached (12 aprox.), a training set is generated with different translations and scales over a uniform grid. Then, a kernel-based regressor is trained for each dimension of the state-space separately using 13 facial features. For each facial feature an elliptical patch of 15 pixels of diameter is extracted, so the representation of a face is the concatenation of those patches, called a facial descriptor. For recognition they use a random-ferns classifier with 40 ferns in 17 levels. When sufficient samples of the track are collected (typically 10), identification is performed marginalizing over the result of random-ferns classifier of the facial descriptors at each frame of the track. In the experiments, up to five people can be tracked at 15fps (multi-core machine at 1.86 GHz), while feature localization can only be computed at 2fps. In general 10 faces are required for recognition, so reliable recognition can then be computed in less than 5s per person. The best results are obtained by taking an episode of TV program, using part of the frames for training, and the rest of them for testing, obtaining $97 \pm 2\%$ of precision with 20% of recall. However, when the system is trained with data from one program and tested in other, the performance decreases significantly.

In contrast to the work of [1], our approach requires only a single frame for training and in average only three frames for recognition, instead of 10. We consider that it is more robust under different environmental conditions, as our tests were done in an indoor environment with natural lighting, while TV programs are recorded in studios with controlled lighting. It seems also that our approach has lower computational requirements as our average recognition time is about half of theirs using similar computers.

3. Scale invariant feature transform

SIFT (Scale invariant feature transform) [6] is a method to extract features invariant to rotation and scaling, and partially invariant to changes in illumination and 3D camera viewpoint. Thus, it seems adequate for face recognition for service robots, where the viewpoint, scale and illumination conditions are variable. The SIFT features are extracted as follows: (i) Detect scale–space extrema, searching over all scales and image locations. Potential interest points invariant to rotation and scale are computed using a differential of a Gaussian. (ii) Localize key points, detecting local extrema and removing low contrast points. (iii) Assign orientations to key points based on the local image gradient. After the points are localized, for each key point a descriptor is obtained by calculating an histogram of local oriented gradients and storing the bins in a 128 dimensional vector. The descriptor also includes the position where the key point is located, the scale where it was found, and the orientation assigned in the previous step.

3.1. Face recognition using SIFT

The use of SIFT for face recognition has been proposed before. A first attempt is the one by Bicego *et al*. [3]. They analyzed 3 different matching strategies. The best results were obtained by subdividing the image using a regular grid. The experiments were realized with the BANCA database considering 52 subjects, each one takes 5 images for training and 7 for testing. All images were resized to 210×200 pixels, equalized and adjusted to fit the eyes with the correct areas for matching.

A second approach is presented by Luo *et al*. [7]. They focus on an automatic process for grouping the key points in the face. For that purpose, they first obtain SIFT key points, and then use a *K-Means* clustering scheme to build k groups of key points. Later, the similarities between two images are calculated using a local similarity measure between corresponding groups, and a global similarity between the whole images. For the experiments they use images from the FERET and CAS-PEAL databases; all images were normalized. The results obtained in the experiments with expression variations report an accuracy of 97% for the clustering method, and 94% for the the grid–based method. However, these results decrease dramatically when illumination variations are included.

Although these two approaches show good results, they were evaluated on face databases with controlled conditions, segmented and normalized; thus are not directly applicable for human–robot interaction. Similarly, we also restrict the SIFT features to certain regions in the face, but these are different as explained below. We also integrate the information form several frames using a Bayesian approach.

4. General overview

Our methodology is divided in three phases:

1) In the first phase frontal faces with variations of ± 20 degrees are detected. In this work we use the rapid object detection scheme based on a boosted cascade of simple feature classifiers [13].

2) Face tracking: After detecting a face, its size (a rectan-

Figure 1. Examples of the performance of the tracking algorithm under different conditions. In each image, the inner rectangle corresponds to the Viola–Jones face detector; and the outer rectangle to the search window.

gular region) is used to define a search window in the next frame, reducing the search space. The search window is defined by increasing each rectangular side by 2/3 of its previous length. If the system is unable to recognize a face in the current window it searches in the whole image. Despite its simplicity this method is extremely robust under different illumination conditions and even with some occlusions. Figure 1 shows some results of the tracking algorithm under different conditions.

3) Face recognition: The mayor contribution of this paper is on the face recognition process once a face has been detected. This phase is described in detail in the following section.

5. Face recognition process

Several steps are followed to recognize faces. These are described in the following sections.

5.1. Image preprocessing

In order to reduce the influence of different illumination conditions in the face recognition process, a preprocessing step is performed on all images. The image is enhanced by equalizing its histogram. This step adjusts the contrast of the image, improving details in the face. Later, a simple compensation of the illumination is implemented as follows [8]. The original face image is divided in 16 regular regions. The average of each region is obtained and a simplified image of 4×4 is created with the average values. This image is bilinear interpolated until obtaining the original image size. Finally, the complement of this image is added to original image to get an image with uniform illumination. This process is applied to all faces in the database and to the images used for face recognition.

5.2. Feature detection

Once a face is detected, our algorithm finds the position of the eyes. We use again the rapid object detection algorithm of Viola and Jones [13] but this time trained to find eyes. When the position of at least one eye is known, we obtain the three regions: a region for the right eye, a region for the left eye and a region for the mouth and nose. The position and size of the nose-mouth region (and of the other eye if necessary) is estimated from the eyes and face regions based on standard face measures. For each of these regions the SIFT features are extracted and used for recognition (see Figure 2).

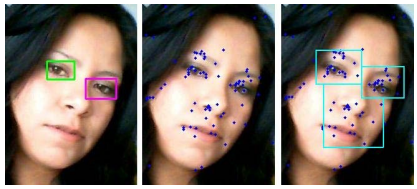

Figure 2. Feature extraction: (a) detection of the eyes, (b) extraction of SIFT points, (c) the regions used for recognition: left eye, right eye, and mouth–nose.

We concentrate on the SIFT points around the eyes, nose and mouth that normally have a high contrast and are easy to identify. The use of only these regions is important because we are able to eliminate SIFT features of probably irrelevant areas, such as hair, clothes or background, and we are also able to improve the matching process as it is performed only between corresponding regions, i.e., SIFT features on the left eye against SIFT features of the stored left eyes, reducing possible miss matches. The total amount of matches found between a new face and a stored face is the sum of the founded matches in the corresponding regions.

5.3. Adding a new face

The system is told when a new face needs to be stored. Once a face is recognized, the SIFT points of its three regions are obtained and stored along with the number of points obtained for that face. The user provides a name for later identification. The total number of recognized SIFT points for each face is stored in a feature vector $\vec{TP} = \{tp_1, tp_2, \ldots, tp_n\}$ for the n faces. A different number of points can be obtained for each face, so also a threshold vector is defined and stored, $\vec{T} = \{t_1, t_2, \ldots, t_n\}$, where each entry is a percentage of the total SIFT points in the face (in the experiments we used 5, 11, 15 and 20%). A single frame is used to obtain the SIFT points for each face.

5.4. Matching strategy

Our matching strategy is based on the criterion used by Lowe [6] using an Euclidean distance but restricted by the three zones, as previously described. Once the three regions are detected in a face and their SIFT points identified, we proceed to see how many matches the new image has with each image stored in the database. At the end of this stage,

a vector of similarities $\vec{S} = \{s_1, s_2, \ldots, s_n\}$ is obtained between the face detected and all the n faces registered in the database. We used three different criteria for successful matching:

1. Condition 1: If $s_i > t_i$ for any i.
2. Condition 2: Condition 1 plus the following condition:

$$max s_i - 2nds_i \geq 2 \times 2nds_i$$

where max is the element with maximum value in $\vec{S}$ and $2nd$ is the second maximum value.

3. Condition 3: Condition 1 plus the following condition.

$$max s_i - avrg(\vec{S}) \geq 2 \times avrg(\vec{S})$$

where $avrg(\vec{(S)})$ is the average of the all the $s_i \in \vec{S}$ except the maximum.

5.5. Recognition process

If a face is not discarded, we use video streaming and a Bayesian approach to improve the recognition rate. The idea is to evaluate the probability of a face f_i given the information from the image features s, that we will denote as $P(f_i|s)$, for each face i stored in memory, and update the probabilities using information from several frames. Using Bayes:

$$P(f_i|s) = \frac{P(s|f_i)P(f_i)}{P(s)} = \frac{P(s|f_i)p(f_i)}{\sum_{k=1}^{n} P(s|f_k)P(f_k)} \quad (1)$$

$P(f_i)$ is initialized as $\frac{1}{n}$ for the n stored faces. $P(f_i)^t$ at instance t is equal to $P(f_i|s)^{t-1}$ at the previous instance $t-1$, thus considering the information from previous frames. $P(s|f_i)$ is estimated from the similarity of the feature vectors. In this paper, we employed two approaches, and *absolute* and a *relative* scheme to obtain these probabilities. The *absolute* probability, takes the percentage of similarities (that is, similar SIFT points s) of each face with respect to the similarities in all the faces:

$$P_{abs}(s|f_i) = \frac{s_i}{\sum_{k=1}^{n} s_k} \quad (2)$$

The relative scheme, takes the percentage of the relative similarities with respect to the total number of recognized points of each faces with respect to the relative similarities of all the faces:

$$P_{rel}(s|f_i) = \frac{\frac{s_i}{tp_i}}{\sum_{k=1}^{n} \frac{s_k}{tp_k}} \quad (3)$$

where tp_i is the total number of SIFT points found for the i-th. face.

All the probabilities $P(f_i|s)$ are stored in a vector that we will be denoted as: $P(F) = \{P(f_1|s), P(f_2|s), \ldots, P(f_n|s)\}$. A person is recognized when its probability is a clear winner over the rest. In this paper, we used the following condition:

$$max P(f_i|s) - 2ndP(f_i|s) \geq 2 \times 2ndP(f_i|s)$$

where max is the element with maximum value in $P(F)$ and $2nd$ is the element with the second highest probability.

If a face is recognized the probabilities are reinitialized to uniform. The same happens if no face is recognized after 10 frames.

6. Experiments

The goal of this face recognition module is to incorporate it into a service robot, where unknown people are expected to be around the robot and the robot can be asked to learn on-line a new face. For these experiments we consider 10 different people. The robot is instructed to register each person, alternating introduction of new persons and recognition. Each person is placed in front of the robot, with a distance of 1 to 5 m, during approximately 100 frames. During the first 50 frames the robot verifies that the person is not already on its database. It then uses a single frame to register the new person (incorporating her/him to those already in the database), and then it tries to recognize her/him in the next 50 frames. We used images of 640×480 pixels in the experiments. It is important to remark that once the image is register in the database, non off-line scaling or image processing are realized for further executions.

We performed experiments with the three matching criteria, with the *absolute* and *relative* probabilities and with four different threshold values, namely: 5%, 11%, 15% and 20% (matches between the SIFT points in memory against SIFT points in the image).

We performed experiments with frames where the person is unknown to the robot. Results for unknown people are shown in Table 1. For instance, using the first criterion (condition 1), *absolute* probabilities (equation 2) and 5% as threshold value (i.e., considering a recognition with 5% or more of matched SIFT points), we have a precision of 92.77%, which means that the system will produce false positives (recognize an unknown person) in roughly 7 out of 100 cases. These results are the average of 166 frames where the system has already identified a face (frames where a face is not detected are discarded) and the person is unknown to the robot.

For known persons, we used 441 frames. Precision results are shown in Table 2 and recall results are given in Table 3. As expected, increasing the threshold value, increases precision but decreases recall. The precision results range between 86.75% up to 99.4% for unknown people, depending on the threshold value and the probability scheme used.

Table 1. Precision results for unknown persons, where *Thrs* is the threshold value, *C1, C2* and *C3* and the three different recognition criteria, and A and R are the *absolute* and *relative* schemes for obtaining the probabilities.

Thrs.	C1-A	C1-R	C2-A	C2-R	C3-A	C3-R
5%	.928	.868	.982	.988	.970	.958
11%	.988	.940	.988	.994	.988	.982
15%	.988	.970	.994	.994	.994	.994
20%	.994	.994	.994	.994	.994	.994

Table 2. Precision results for known persons (see table 1 for notation).

Thrs.	C1-A	C1-R	C2-A	C2-R	C3-A	C3-R
5%	.967	.983	.974	.994	.978	.985
11%	.989	1.00	.993	1.00	.994	.994
15%	.988	.994	.993	1.00	.994	.993
20%	1.00	1.00	1.00	1.00	1.00	1.00

Table 3. Recall results for known persons (see table 1 for notation).

Thrs.	C1-A	C1-R	C2-A	C2-R	C3-A	C3-R
5%	.573	.550	.463	.402	.537	.498
11%	.440	.419	.366	.326	.427	.400
15%	.389	.376	.324	.292	.378	.359
20%	.335	.324	.285	.260	.331	.319

For known people the precision and recall ranges vary between 100% of precision with 33% of recall to 57.32% of recall with 96.65% of precision. In general, the *absolute* scheme used to evaluate probabilities tend to perform better than the *relative* scheme. The system gives very few false positives in both cases, where a person is already known or for new unknown faces. Also, as expected, depending on the conditions used to recognize a person we can increase accuracy but decrease recall or vice verse. The second condition used for recognition has higher precision but smaller recall.

Given that the system is for human–robot interaction, it is not necessary that the system recognizes the person in every frame, so the recall is not critical. Given this, the results are very good, as we can approach a 100% precision with a relatively good recall. Also, the system in general recognizes when the person is not in the database, as the precision for unknown persons is over 90% for almost all the experiments.

The previous results show the behavior of the system while it is learning new faces on-line. After completing a database with 10 faces, we ran again the experiments with similar results. The average time per frame for the whole process takes about 1.2 seconds, and the average number of frames required to take a decision is 2.5, which means that

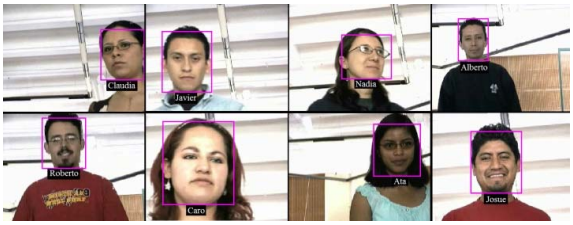

Figure 3. Some results produce by the system with people identified by the robot under different conditions.

Figure 4. Several images of a person in Yale's database.

the system can identify a person in approximately three seconds in average, running on a Pentium D at 2.8 GHz with one Gb of RAM. Some results of the execution of the system are shown in Figure 3. As can be seen from the results, successful recognitions are achieved with different face orientations and distance conditions.

We also tested the system with Yale's face database [4] which has images of 28 persons under different illumination conditions. An example of different images of one person is shown in figure 4. For this experiment we trained our system with one image from each person, and tested with the other images. We tested for one, two, ... up to 28 known persons for the system, and the other persons as unknown. For each case, we used 1, 5, 10, 15 and 20 frames of a person for recognition with our Bayesian approach. These frames have different illuminations so this make more difficult the recognition process. The results are summarized in figures 5 and 6 that show the precision and recall vs. the number of persons, respectively. Each experiment was repeated 10 times with a random selection of the images, the graphs show the average performance. As expected, the precision declines as the number of persons increases, but it is still about 85% for 28 persons when using 20 images. We can also observe that combining several images has a significant impact on precision and recall, although there is no much difference above 10 frames. The system has in general a high precision with not as good recall, which for robot–human interaction is a good compromise.

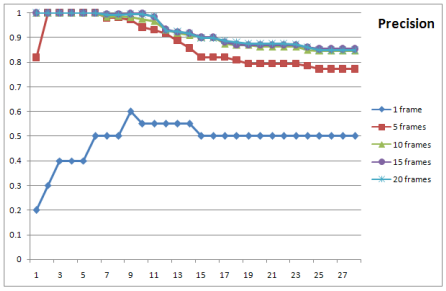

Figure 5. Precision results for the Yale's database.

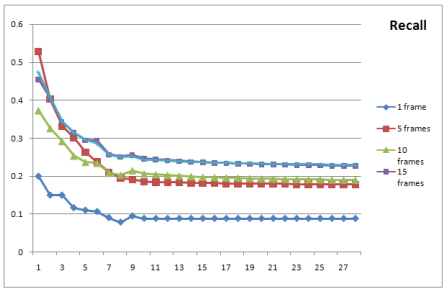

Figure 6. Recall results for the Yale's database.

7. Conclusions and future work

We have developed a system able to learn on-line faces for later recognition in a service robot framework. The system can work under different illumination conditions and in unstructured environments with very promising results. A new face can be learned on-line from a single frame and then recognized the next day under different conditions. We used SIFT features extracted from the face for the recognition process with two important enhancements: (i) we extract features only on selected regions of the face and perform the matching process only between corresponding regions, and (ii) we combine information from several frames based on a Bayesian approach. Both strategies produce very competitive results against related work that assumes more restricted conditions. Depending on the application domain, the user can define different conditions to increase the accuracy or recall. In our experiments, a 5% threshold value provides a good compromise between precision and recall. It is also shown from the experiments, that the proposed system produces very few false positives.

At the moment, we assume that the person to be recognized gets in front of the robot, although the system has some tolerance in terms of the orientation and distance to the face ($\pm 20^o$ and between 0.5 m and 5 m). As future work, we are coupling our scheme with a navigation algorithm to move the robot in front of a person.

Acknowledgments

We thank Professor David Lowe for providing us his code for SIFT. The Extended Yale Face Database B was used for some of the experiments. This research was supported in part by CONACYT under project No. 47968.

References

[1] N. Apostoloff and A. Zisserman. Who are you? - real time person identification. In *Proc. British Machine Vision Conference*, pages 509–518. BMVA, 2007.

[2] O. Arandjelovic and A. Zisserman. Automatic face recognition for film character retrieval in feature-length films. In *Proc. Computer Vision and Pattern Recognition*, pages 860–867, 2005.

[3] M. Bicego, A. Lagorio, E. Grosso, , and M. Tistarelli. On the use of sift features for face authentication. In *Proc. Computer Vision and Pattern Recognition Workshop*, page 35, 2006.

[4] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 23(6):643–660, 2001.

[5] C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Trans. Image Processing*, 11(4):467–476, 2002.

[6] D. Lowe. Object recognition from local scale-invariant. In *Intl. Conf. on Computer Vision*, pages (2):1150–1157, 1999.

[7] J. Luo, Y. Ma, E. Takikawa, S. Lao, M. Kawade, and B. L. Lu. Person-specific sift features for face recognition. In *Intl. Conf. on Acoustic, Speech and Signal Processing (2)*, pages 593–596, 2007.

[8] G. Ramírez-García. *Detección de Rostos con Aprendizaje Automático*. Master thesis, INAOE, Puebla, Mexico, 2006.

[9] J. Sivic, M. Everingham, and A. Zisserman. Person spotting: video shot retrieval for face sets. In *Proc. Intl. Conf. on Image and Video Retrieval*, pages 226–236, 2005.

[10] J. Sivic, C. L. Zitnick, and R. Szeliski. Finding people in repeated shots of the same scene. In *British Machine Vision Conference*, pages 909–918, 2006.

[11] K. Song and W. Chen. Face recognition and tracking for human-robot interaction. In *IEEE Intl. Conf. on Systems, Man and Cybernetics (3)*, pages 2877–2882, 2004.

[12] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[13] P. Viola and M. Jones. Rapid object detection using boosted cascade simple features. In *Proc. Computer Vision and Pattern Recognition*, pages 511–518, 2001.