

Face Recognition for Mobile Robots

Robin Hewitt
rhewitt@acm.org

Abstract

A method for detecting and recognizing faces from one input image was devised and implemented. This method is intended to serve as the initial phase in a face-learning system for a personal or social robot. The method described here uses a cascaded search with haarlike features to quickly eliminate image regions that do not match the input model. The remaining regions are evaluated at three scales using image-gradient features. Closeness to the input model is evaluated as the sum of local deformations to bring the search region and the input model into correspondence. The resulting method is computationally efficient and was able to recognize the original face under moderate changes in lighting, pose, and expression with a true-positive rate of 85% when the threshold for per-frame false-positive rate was adjusted to be 15%.

1. Introduction

Implementing face recognition in the context of a social or personal robot presents unique challenges. While there are now many excellent methods for learning objects and object classes from database images, the applicability of these learning methods to the personal-robot context is limited. Face databases often contain biases that differ from the population bias in the robotics context. A case in point is camera position. Images in face databases are typically captured at eye level. But for both psychological and economic reasons, social robots are typically short and have their camera positioned well below eye level. Faces present a significantly different appearance when viewed from this angle. Other relevant biases common to most face databases include frontal lighting, pale skin, and a lack of motion blur.

The physical configuration for a social robot can vary widely, as can the camera used. Offsetting these confounding variables, the number of faces that the robot might be expected to recognize is relatively small – from one to perhaps twenty.

The goals in a robotics deployment context are different as well. When face recognition methods are deployed for airport security or for mugshot identification, the concern is to improve the odds of catching a criminal. In these situations, the success criterion is itself a statistical measure.

The commercial success of a personal robot, however, may be better supported by the ability to learn to recognize just a few faces, and to do so well in every deployment. In this case, the deployment context is open ended, the recognition criterion is no longer a statistical measure, and success now depends on the ability to adapt.

The personal-robot context includes benefits as well as challenges. An adaptable recognition algorithm can take advantage of user willingness to provide feedback. For users to remain willing to provide input to a learning algorithm, they should see that their input is effective, in other words, that the robot learns. In addition, the interface should also be as simple and easy to use as possible. For face recognition, a reasonable user-input scenario is to capture one video frame and add minimal markup. If face recognition from this initial input is good enough to keep the user interested, they will continue to interact with the robot and be willing to continue on to a next phase of learning. A related benefit in this context is that, as the robot performs face recognition, it is in an excellent position to acquire additional, useful data for future learning.

In addition to good recognition performance, the initial algorithm should run easily in real time. Lowe’s well-known recognition approach, described in [5] offers a one-step learning paradigm. This method has been shown to be robust to many types of variations as well. It is, however, computationally demanding compared to statistically learned models such as the cascaded Haarlike filters used by Viola and Jones in [10].

This paper presents a method for combining the efficiency of a cascaded search with the ease-of-use of a single-input learning paradigm. It utilizes elements from both Lowe’s approach and the methods introduced by Viola and Jones. The goodness of a match is determined by thresholding against a deformation-based distance measure.

2. Related Work

Viola and Jones [10] demonstrated an efficient face detector based on brightness differences between adjacent rectangular regions. These Haar-wavelet-like features can be located very rapidly by building an integral image.

Balas and Sinha [1] extended this feature type by lifting the requirement that the rectangular regions be adja-

cent. They point out that any two rectangular regions can be paired. The regions are represented by their average pixel brightness, and the difference in brightness levels used as a descriptor. They call these features dissociated dipole descriptors. They share the same efficiency benefit as the haar-like descriptors used by Viola and Jones, but have greater expressive power and can be made less noise sensitive by eliminating noisy transition regions between light and dark image regions. The authors use these to learn a descriptor pool which in turn can be used for object representation.

In the approach presented here, these dipole descriptors are used in a different way. Because each pole of a dipole can be made smaller than the bright or dark region it captures, these descriptors can represent these large-scale features with soft localization. Since they're not constrained to be adjacent, these large, high-contrast regions can be individually selected and paired to create robust, high-level appearance filters that efficiently remove all but a small fraction of the image from the search space.

This work also builds on Lowe [5], and extends it in several ways. First, it combines gradient-based features with haar-like ones to improve search efficiency. Search speed is also increased by decoupling feature searching from feature learning. By searching directly for the gradient-based features in an efficient manner, the Laplacian pyramid that's used to identify interest points is eliminated during search. Additionally, multiple types of interest-point detectors can be used to enrich the object description without incurring a search-time performance penalty.

3. Method

The initial input into this system is a single image. A user presents a frontal-view face example as a video-frame capture and draws a bounding rectangle to indicate its location. Users are encouraged to include enough margin to capture areas of the head that are likely to be distinctive. For example, profile lines at the top and sides of the head can be distinctive features, even against a cluttered background [3]. The chin line can also be distinctive, as can the shadowed area beneath the chin.

Filtering by large-scale-contrast features Some knowledge of face shape is used at this point to help select good regions for the large-scale dipole features. The forehead, when visible, nose, and cheeks are generally brighter than the eye region, the shadowed area under the nose, and the area under the chin. Currently users are asked to mark the centers of both eyes and the nose tip as an aid to localizing these areas. Eventually, this process is expected to be fully automated.

Not more than three levels are used for the cascade. Each level contains three strong dipoles. A dipole feature is con-

sidered present if the difference in average brightness between the light and dark regions is half the observed level in the input image. During search, any region that passes this test for two of the three dipole regions at that cascade level proceeds to the next level. The 1/2 threshold level was selected based on an ROC-curve analysis in which threshold level was varied from 40 to 100% of the original brightness difference. Below 60% of the original difference, there were no false negatives, provided the dipole areas were well within each region. To ensure that this condition is met, the rectangles are brought out to the edges of their intensity zone, then the width and height are reduced by 25%. The edge of the zone is detected based on $\frac{dI}{dA}$, the rate of change in average brightness with respect to area. This is similar in spirit to the approach used in [6] to detect maximally stable extremal regions.

During search, potential face regions at three scales and at each image location are evaluated with the large-scale-feature cascade. This cascade is implemented using the integral-histogram method described in [10] and [1]. It thus runs very quickly. The output from the cascade is a list of pixels, with each pixel associated with an image region at one scale. The number of surviving pixels from a typical 320 x 240 image ranges from few hundred to a few dozen for each scale. Since the cascade features use soft localization, image pixels that pass through the cascade form compact clusters in the image. These are combined using region growing, further reducing the total number of potential face regions. By the next stage of the search process, the total number of search regions has been reduced from 3 x 70,000 (scales x pixels) to about 10 or 20.

Region selection Once the image has been filtered against the large-scale dipole patterns, the remaining candidate positions are evaluated using small-scale descriptors with more precise localization. The small-scale descriptors are similar to the SIFT descriptors presented in [5] or the HOG descriptors presented in [3]. They consist of a 4x4 cell array of partially overlapping subregions. Each subregion is represented as a weighted histogram of eight gradient directions. Partially overlapping descriptors were found in [3] to give better discrimination than the non-overlapping SIFT descriptors. In both [5] and [3], the histogram contribution from each pixel's gradient direction is weighted by the gradient magnitude at that location. However, sensitivity to lighting changes can be significantly reduced by scaling the gradient-magnitude weighting by the local brightness as $w = \frac{\|G\|}{\bar{I}}$, where \bar{I} can be obtained by various averaging methods, such as gaussian filtering. The method and descriptors presented here were evaluated using both weighting methods. Using brightness-normalized weighting noticeably improved performance. Average brightness for this purpose was obtained by mean-filtering each 3x3

pixel neighborhood, using the integral image that had been set up for the cascade.

When building the model, interest-points for each descriptor are located using both Difference of Gaussians, as described in [5], and the Harris Corner detector, described in [4]. Each small-scale descriptor is thus associated with a relative (x,y) location in the model. During search, however, these interest-point operators are not applied directly. Instead, each descriptor is compared directly to the image gradients within a local region surrounding the place it would be expected to occur relative to each candidate position and scale that survived the preceding cascade.

One advantage of this approach is that several complementary interest-point detectors can be used to generate a richer object description and to give good coverage of the object’s appearance surface. During search, the region in which each feature can occur is greatly reduced after the cascade, such that it’s computationally efficient to directly apply each descriptor over the remaining area.

The metric by which candidate regions are evaluated for model similarity requires explanation. The standard deviation for descriptor similarity between a feature in the model and that feature’s appearance in the wild is unknown. Without more knowledge about appearance variability, the absolute value of a descriptor’s similarity metric may be a poor guide to the quality of the match. Instead, each feature is assigned the (x,y) location corresponding to its best match within the local search area. Disparity with the model is then computed as the sum of L1 distances, in the x and y dimensions, between where each feature occurred in the model and where its best local match was found within the search region. Intuitively, this approximates the cumulative local strain placed upon the model as each feature location is pulled and stretched to align with its best match in the image region.

4. Data and Testing

Five face images from each of five individuals were captured using the robot’s webcam. One of these was used as an input image and the rest for testing. In addition, fifty background images were captured from three home environments.

Each of the five face models was tested against 1) the remaining four examples of that person’s face, 2) all background images, 3) the face images from all other individuals, and 4) the combined negative-examples set of backgrounds plus face images from all other individuals.

5. Results and Discussion

Faces of the same individual that were a good match to one of the three search scales were typically less than a distance

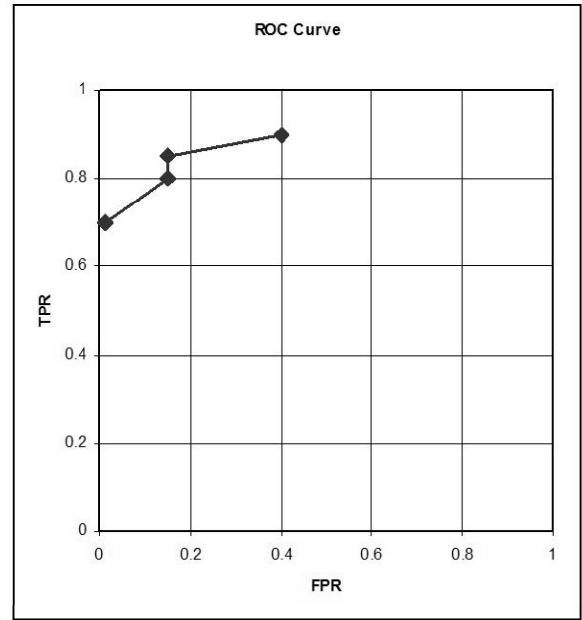


Figure 1: ROC curve for all test data. Positives and negatives are on a per-frame basis. The negative-examples test set for each person included face images from each of the other individuals as well as all 50 background images.

of 4.2 to 4.5 away from the input model. The average distance for a background image from an input model was 5.2. Search images that contained a different person’s face had an average distance of 5.5 from the input model. It’s notable that faces of different individuals were no more likely to trigger a false positive than were the background images. The reason for this behavior was that the large-scale dipole features captured enough of each individual’s face biometrics to prevent matches between individuals in most cases.

Setting the threshold for matching at 4.5 gave a true-positive rate of 85%. At that threshold level, 15% of the non-same-face images (combined background and other faces) registered as false-positives. The ROC curve for the combined dataset is shown in Figure 1. The percentage results (TPR and FPR) are on a per-frame, not on a per-pixel, basis.

The goal for this algorithm is to maintain user interest as learning continues. An 85% recognition rate should be more than adequate to show that the robot has been responsive to the user’s initial input. In addition, after users have given the robot this initial training, they will be interested in how well it works. It’s thus reasonable to assume that many, and perhaps most, of the frames immediately following the initial training will contain positive, rather than negative examples. Thus the usual concern that negative examples are

extremely rare, is not initially a concern in this deployment context. By the time images of the user's face become a rare event, the robot will have had opportunity to acquire additional information. In particular, by tracking the user's face, the robot can begin immediately to gather additional positive examples. My next effort for this recognition approach will focus on incorporating motion-based segmentation and tracking.

All results for searches against a same-face image are shown in Figure 2. Figure 3 shows examples of the background images.

References

- [1] B. Balas and P. Sinha, "Receptive field structures for recognition," *Technical Report, Computer Science and Artificial Intelligence Laboratory, MIT*, MIT-CSAIL-TR-2005-015 (AIM-2005-006, CBCL-246), March 1, 2005, <http://publications.csail.mit.edu/tmp/MIT-CSAIL-TR-2005-015.pdf>.
- [2] P.J. Burt and E.H. Adelson, "The Laplacian Pyramid as a Compact Image Code," *IEEE Trans. on Communications*, Vol. 31, pp. 1532-1540, 1983.
- [3] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *CVPR 2005*, Vol. 1, pp. 886-893, 2005.
- [4] C. Harris, M. Stephens, "A Combined Corner and Edge Detector," *Alvey Vision Conference*, pp. 147-151, 1988.
- [5] D. Lowe, "Object Recognition from Local Scale-Invariant Features," *Proc. International Conference on Computer Vision*, (ICCV1999) Vol. 2, pp. 1150-1157, 1999.
- [6] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust Wide-Baseline Stereo from Maximally Stable Extremal Regions," *Image and Vision Computing*, 22(10):761-767, 2004.
- [7] F. Provost and T. Fawcett, "Robust Classification for Imprecise Environments," *Machine Learning*, 42(3):203-231, 2001.
- [8] J. Sivic and A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos," *Proc. International Conference on Computer Vision (ICCV2003)*, Vol. 2, 2003.
- [9] J. Sivic, F. Schaffalitzky, and A. Zisserman, "Object Level Grouping for Video Shots," *Proc. 8th European Conference on Computer Vision (ECCV 2004)*, Vol. 2, 2004.
- [10] P. Viola and M. Jones, "Robust Real-time Object Detection," *Proc. IEEE Int. Workshop on Statistical and Computational Theories of Vision*, Vol. 2, pp. 1150-1157, 2001.



Figure 2: All face images with best-fit region and distance from input image.

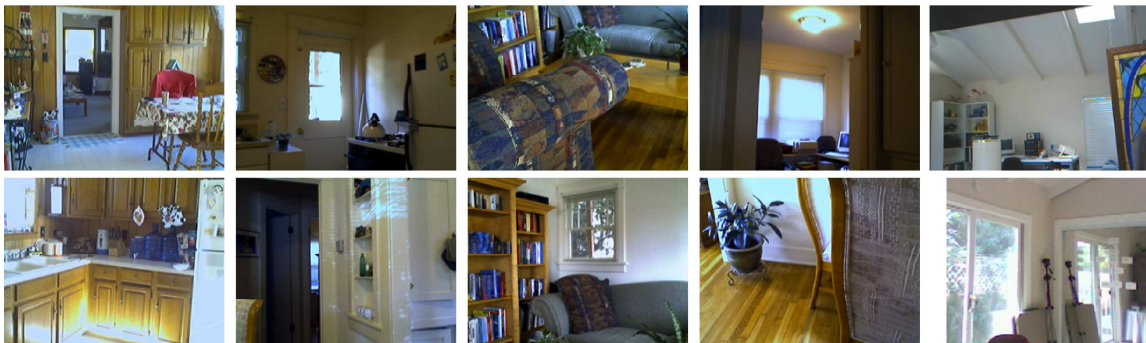


Figure 3: Example background images.