Face recognition using Histograms of Oriented Gradients

O. Déniz,⇑ G. Bueno, J. Salido, F. De la Torre

Universidad de Castilla-La Mancha, E.T.S. Ingenieros Industriales, Avda. Camilo Jose Cela s/n, 13071 Ciudad Real, Spain
Carnegie Mellon University, Robotics Institute, 211 Smith Hall, 5000 Forbes Ave., Pittsburgh, PA 15213, USA

Article info
Article history:
Received 24 September 2009
Available online 20 January 2011
Communicated by H. Wechsler

Keywords:
Face recognition
Histograms of Oriented Gradients
Active Appearance Models

Abstract
Face recognition has been a long standing problem in computer vision. Recently, Histograms of Oriented Gradients (HOGs) have proven to be an effective descriptor for object recognition in general and face recognition in particular. In this paper, we investigate a simple but powerful approach to make robust use of HOG features for face recognition. The three main contributions of this work are: First, in order to compensate for errors in facial feature detection due to occlusions, pose and illumination changes, we propose to extract HOG descriptors from a regular grid. Second, fusion of HOG descriptors at different scales allows to capture important structure for face recognition. Third, we identify the necessity of performing dimensionality reduction to remove noise and make the classification process less prone to overfitting. This is particularly important if HOG features are extracted from overlapping cells. Finally, experimental results on four databases illustrate the benefits of our approach.

1. Introduction

Face recognition has been a long standing problem in computer vision. It has recently attracted significant attention due to the accessibility of inexpensive digital cameras and computers, and its applications in biometrics and surveillance (see Zhao et al., 2003; Chellappa et al., 1995; Samal and Iyengar, 1992; Chellappa and Zhao, 2005) for recent surveys of face recognition).

Central to the success of face recognition are the feature representation and the classification method. In this paper, we will focus on the former. Broadly speaking, we could classify the features for face recognition as geometric or photometric (view based). The latter seem to have prevailed in the literature (Zhao et al., 2003). There exist a large number of features, starting from the influential Eigenfaces (Principal Component Analysis) (Turk and Pentland, 1991), Gabor wavelets (Amin and Yan, 2009), Local Binary Patterns (Ahonen et al., 2004), Error-Correcting Output Codes (Kittler et al., 2001) and Independent Component Analysis (ICA) (Bartlett et al., 2002) among others.

Histograms of Oriented Gradients (HOGs) (Lowe, 2004) are image descriptors invariant to 2D rotation which have been used in many different problems in computer vision, such as pedestrian detection (Bertozzi et al., 2007; Wang and Lien, 2007; Chung et al., 2008; Watanabe et al., 2009; Baranda et al., 2008; He et al., 2008; Kobayashi et al., 2008; Suard et al., 2006; Zhu et al., 2006; Perdersoli et al., 2007a,b). Recently, in (Albiol et al., 2008) the authors successfully applied HOG descriptors to the problem of face recognition. In that work, a set of 25 facial landmarks were first localized using the Elastic Bunch Graph Matching framework (see Wiskott et al., 1997). The HOG features extracted from the vicinity in each of these 25 facial landmarks were used for classification, using nearest neighbor and Euclidean distance. In this paper, following (Albiol et al., 2008), we further explore the representational power of HOG features for face recognition, and propose a simple but powerful approach to build robust HOG descriptors. In particular, three are the main novelties: (1) build the HOG descriptor using a regular grid, (2) build and combine the HOG descriptors at different scales, (3) apply a linear dimensionality reduction to remove noise, make the classifier more efficient (i.e. reduce dimensionality) and less prone to overfitting. Our results in four standard face databases support the proposed method.

The rest of the paper is organized as follows. Section 2 describes HOG in detail, as well as our approach. In Section 3 we describe the experimental validation. Section 4 finalizes the paper with the main conclusions.

2. Building a representation for face recognition using HOGs

The algorithm for extracting HOGs (see Dalal and Triggs, 2005; Lowe, 2004) counts occurrences of edge orientations in a local neighborhood of an image. In our case, the image is first divided into small connected regions, called cells, and for each...
cell a histogram of edge orientations is computed. The histogram channels are evenly spread over $0–180^\circ$ or $0–360^\circ$, depending on whether the gradient is ‘unsigned’ or ‘signed’. The histogram counts are normalized to compensate for illumination. This can be done by accumulating a measure of local histogram energy over the somewhat larger connected regions and using the results to normalize all cells in the block. The combination of these histograms represents the final HOG descriptor. Invariance to scale and rotation may be also achieved by extracting descriptors from only salient points (keypoints) in the scale space of the image following a rotation normalization. The steps involved are:

1. Scale-space extrema detection.
2. Orientation assignment.
3. Descriptor extraction.

The first step is intended to achieve scale invariance. The second step finds the dominant gradient orientation. All the orientation counts are then made relative to this dominant direction. Fig. 1 shows an example patch with its corresponding HOGs.

In (Albiol et al., 2008) (and the shorter version Monzo et al., 2008) the authors successfully applied HOG descriptors to the problem of face recognition. In that work, faces were previously normalized in scale and orientation, so the steps for scale-space extrema detection and orientation were not necessary. A set of 25 facial landmarks were localized using the Elastic Bunch Graph Matching framework (see Wiskott et al., 1997) with HOG features. The HOG features extracted from the vicinity of each of these 25 facial landmarks were used for classification, using nearest neighbor and Euclidean distance. It is important to note that for a new testing face, the matching procedure makes use of the eye’s position (the eyes were assumed to be at a fixed position after normalization).

A potential drawback of the approach taken in (Albiol et al., 2008) is that the final error may crucially depend on the reliability of the landmark localizations. Our hypothesis in this paper is that such approach may not work well when landmarks are not precisely localized due to occlusions, strong illuminations or pose changes. Thus, in this work we propose to first normalize the face and then extract HOG features from a regular grid. The grid is formed by placing equal side patches around a first cell centered in the image, until the whole image is covered.

On the other hand, the size of the patch used to extract the HOG features is important. In (Albiol et al., 2008) the best size for the patch was estimated via cross-validation in the Yale database, prior to using the FERET database for the final experiments. The locality of the extracted features is determined by the patch size. We hypothesize that a better result could be obtained by combining information from different patch sizes. The fusion strategy that will be considered here is the product combination of the classifiers at patch sizes. Note that this combination rule is not optimal since it assumes independence of the classifiers, though empirically has

### Table 1

<table>
<thead>
<tr>
<th>Name</th>
<th>Classes</th>
<th>Total samples</th>
<th>Samples per class (min/avg/max)</th>
<th>Variations present</th>
</tr>
</thead>
<tbody>
<tr>
<td>FERET</td>
<td>1195</td>
<td>3540</td>
<td>2/2.9/32</td>
<td>Facial expression, aging of subjects, illumination</td>
</tr>
<tr>
<td>MPIE-2</td>
<td>337</td>
<td>2509</td>
<td>2/7.4/11</td>
<td>Expression, session</td>
</tr>
<tr>
<td>AR</td>
<td>132</td>
<td>3236</td>
<td>13/24.5/26</td>
<td>Expression, illumination, occlusions (sunglasses and scarves)</td>
</tr>
<tr>
<td>Yale</td>
<td>15</td>
<td>165</td>
<td>11</td>
<td>Expression, illumination, glasses</td>
</tr>
</tbody>
</table>

Fig. 1. Example HOG descriptors, patch size=8 × 8. Each cell of the patch shows the orientation of the gradients.

Fig. 2. Left: Initialized landmarks (red boxes) and result for the AAM fitting in an image from the Yale database. Center: extracted HOG descriptors, patch size = 24 × 24. Right: extracted HOG descriptors, patch size = 64 × 64. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
performed well in our experiments. Suppose \( R \) individual classifiers \( c_k \) \((k = 1, \ldots, R)\) each one trained using HOG features are extracted from different patch sizes. Each classifier assigns one input sample (represented as \( x_k \)) to a label \( L_k = w_1, \ldots, w_m \).

Assume the classifier \( c_k \) gives every output a measurement which is represented as a posterior probability vector, \( P_k = \{p(w_1|x_k), \ldots, p(w_m|x_k)\} \), where \( p(w_j|x_k) \) denotes the probability that the classifier considers that \( x \) was labeled with \( w_j \). The product rule consists of fusing the final decision as:

\[
j = \arg \max_i \prod_{k=1}^{R} p(w_i|x_k)
\]

Note that when several overlapping patches are used, the final feature representation will be highly redundant and if the classifier does not have any mechanism for feature selection it might severely suffer from overfitting. Observe that the human face displays a structure common to all individuals. This implies that some gradient orientations would be very frequent in some specific zones of the face. Other orientations, on the contrary, would never or almost never appear in a given region. This reinforces the idea of the need for dimensionality reduction techniques.

### 3. Experiments

This section describes two experiments. The first experiment evaluates the impact of facial landmark localization in the face recognition performance. The second experiment evaluates the impact of extracting the HOG features from a regular grid and at multiple scales.

---

1 Other evaluated fusion rules were majority voting, mean, median and maximum.
The experimental results were evaluated using four face databases: FERET (Phillips et al., 2000), AR (Martinez and Benavente, 1998), CMU Multi-PIE (Sim et al., 2001) and Yale (Yale face database, 2009). All these data sets cover a wide range of variations and scenarios for face recognition, see Table 1. All the images were previously normalized to 58×50 pixels.

3.1. Experiment 1: robustness to facial feature location

In the first experiment we evaluated the effect of the facial feature localization error on the final recognition performance. We compared the recognition performance using a holistic representation versus a HOG-based representation on facial landmarks. Forty-nine facial landmarks were localized using Active Appearance Models (AAMs) (De la Torre et al., 2007). Fig. 2 on the left shows the initialized set of landmarks and the landmarks found by the AAM in an image from the Yale database. Fig. 2 on the center and right illustrates the spatial domain for the HOG descriptors.

Fig. 3 shows the classification accuracy when using an holistic gray level representation with PCA or LDA (dotted lines), or when automatically detecting the landmarks with AMM and extracting HOG features (continuous line). When we consider absolute recognition performances, we see that the recognition rates when using the HOG representations extracted from the detected facial landmarks is lower when compared to the holistic PCA and LDA (except for FERET). The AAM was not trained under strong illumination and pose changes, and cannot deal with strong occlusions. This leads to large errors in the localization of facial features biasing the recognition performance. Observe that in the FERET database the classification performance of the landmark-based method is better because the landmark location is more accurate (no occlusion, and strong illumination changes). To check this hypothesis we computed a measure of landmark localization dispersion, see Fig. 4. The FERET database has the lower variance in the detected landmarks, whereas the AR has the highest variance (the AR database is the only one that includes major occlusions, like sunglasses and scarves, see Fig. 5). In this case, the precision of landmark detection correlates with the recognition performance using HOG.

Fig. 5. Bad landmark localization examples in the AR database. The green crosses denote the initial search points. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

![Fig. 5](image)

We compared the recognition performance using a holistic representation versus a HOG-based representation on facial landmarks. Forty-nine facial landmarks were localized using Active Appearance Models (AAMs) (De la Torre et al., 2007). Fig. 2 on the left shows the initialized set of landmarks and the landmarks found by the AAM in an image from the Yale database. Fig. 2 on the center and right illustrates the spatial domain for the HOG descriptors.

Fig. 3 shows the classification accuracy when using an holistic gray level representation with PCA or LDA (dotted lines), or when automatically detecting the landmarks with AMM and extracting HOG features (continuous line). When we consider absolute recognition performances, we see that the recognition rates when using the HOG representations extracted from the detected facial landmarks is lower when compared to the holistic PCA and LDA (except for FERET). The AAM was not trained under strong illumination and pose changes, and cannot deal with strong occlusions. This leads to large errors in the localization of facial features biasing the recognition performance. Observe that in the FERET database the classification performance of the landmark-based method is better because the landmark location is more accurate (no occlusion, and strong illumination changes). To check this hypothesis we computed a measure of landmark localization dispersion, see Fig. 4. The FERET database has the lower variance in the detected landmarks, whereas the AR has the highest variance (the AR database is the only one that includes major occlusions, like sunglasses and scarves, see Fig. 5). In this case, the precision of landmark detection correlates with the recognition performance using HOG.

Fig. 6. Classification performance in four face databases. We show the mean and standard deviation over 10 runs, using 50% of the data for training and 50% testing.
The contributions are threefold: (1) to provide robustness to facial feature detection, we propose to uniformly sampling the HOG features, (2) to remove redundancy in the data, improve computational efficiency and avoid overfitting, we propose to use dimensionality reduction in the HOG representation, (3) we show that a decision-level combination of results using HOG features extracted from different image patch sizes significantly improves on choosing a single best patch size. Taking into account these considerations we were able to obtain a significant increase (up to 13%) in recognition performance on the standard FERET database.

Acknowledgments

This work was partially funded by project PI12109-0043-3364 of Castilla-La Mancha Regional Government and the Spanish Ministry of Science and Innovation Funds (project TIN2008-06068).

References


