On using periocular biometric for gender classification in the wild

Modesto Castrillón-Santana,*, Javier Lorenzo-Navarro, Enrique Ramón-Balmaseda

SIANI, Universidad de Las Palmas de Gran Canaria, Spain

ABSTRACT

Gender information may serve to automatically modulate interaction to the user needs, among other applications. Within the Computer Vision community, gender classification (GC) has mainly been accomplished with the facial pattern. Periocular biometrics has recently attracted researchers attention with successful results in the context of identity recognition. But, there is a lack of experimental evaluation of the periocular pattern for GC in the wild. The aim of this paper is to study the performance of this specific facial area in the currently most challenging large dataset for the problem. As expected, the achieved results are slightly worse, roughly 8 percentage points lower, than those obtained by state-of-the-art facial GC, but they suggest the validity of the periocular area particularly in difficult scenarios where the whole face is not visible, or has been altered. A final experiment combines in a multi-scale approach features extracted from the periocular, face and head and shoulders areas, fusing them in a two stage ensemble of classifiers. The accuracy reported beats any previous results on the difficult The Images of Groups dataset, reaching 92.46%, with a GC error reduction of almost 20% compared to the best face based GC results in the literature.

1. Introduction

Humans exhibit an extraordinary ability to extract information from facial images, that serves for multiple purposes in daily human interaction. The reliability of the human system has attracted the attention of psychologists for decades. More recently, Computer Vision researchers are interested in developing automatic systems with similar skills. In this work, we specifically focus on Gender Classification (GC).

Different biometric traits have been applied to the GC problem as gait, hair or clothes, as described in [6, 7, 22, 17]; and [42]. However, most automatic GC systems have been and are being designed to determine the human gender from the facial pattern, as evidenced in the recent NIST survey by [30]. The face is indeed a valid cue for humans, but a deeper analysis of the human system highlights the main importance of the ocular and mouth areas, as suggested by [20]. In this paper, we analyze the significance of the ocular area for this particular problem, but in the wild.

*Corresponding author: Tel.: +34-928-458-743; fax: +34-928-458-711; e-mail: modesto.castrillon@ulpgc.es (Modesto Castrillón-Santana)
The recent survey on ocular biometrics by [31] reviews the results achieved making use of three different modalities: iris, periocular and retina. Among them, we refer below to the first two, the external ones. The human iris is certainly a powerful biometric trait as proved in identity recognition. However, to be reliable, iris based systems require user cooperation and some specific acquisition conditions related to the pattern resolution. A less restricted alternative in terms of user cooperation, is the periocular area, i.e. the eye and its surrounding area. There is an increasing interest in the still relatively unexplored field of periocular biometrics, with recent promising and successful results in identity recognition.

The mentioned survey by [31] also describes the advantages of modalities fusion. Indeed, the fusion of facial and periocular information improves identity recognition performance, proving the benefits in surveillance scenarios, where noise and non-cooperation excludes the use of other ocular cues. Our hypothesis is that the same behavior will be observed in GC even in the wild. Thus, we first study automatic GC based on only the periocular region. This is done analyzing the performance of different local descriptors achieved in the currently hardest dataset. Once compared to face based approaches we evaluate the fusion of both modalities: periocular and face.

The contributions of the work are three-fold: 1) an exhaustive evaluation of local descriptors in the currently most challenging dataset for the GC problem, proving that the periocular is a valid cue for large populations; 2) different descriptors provide complementary information, that serves to setup a score level fusion of selected descriptors to improve the overall periocular based GC performance; and 3) the confirmation that the multi-scale combination of selected features extracted from the periocular, face, and head and shoulders areas steps forward a more accurate gender recognizer.

2. State of the art

2.1. Facial GC

The 2015 NIST evaluation by [30] remarks the differences between constrained and unconstrained or in the wild GC. The best evaluated system was able to reach an accuracy up to 96.5% in a constrained dataset containing around one million samples.

Fig. 1. Sample images respectively of The images of Groups ([18]), The Labeled Faces in the Wild ([21]) and MORPH ([36]) datasets. Their respective original resolutions are 391 × 293, 249 × 249 and 200 × 240 pixels.
However, those accuracies dropped significantly for some particular in the wild datasets. Two remarkable smaller datasets were analyzed: *The Labeled Faces in the Wild* (LFW) by [21], and *The images of Groups* (GROUPS) by [18]; reporting a similar accuracy for LFW, while for GROUPS the best commercial solution dropped to an accuracy of 90.4%. This effect is also confirmed by the research community. Table 1 summarizes recent results on three large datasets, adding MORPH, by [36], to the mentioned GROUPS and LFW. A sample of images of each dataset is presented in Figure 1. Observing Table 1, GROUPS achieves an accuracy that hardly reaches 90%. Compared to the other datasets, there is an evident gap between the accuracy achieved for GROUPS and the rest, being almost 7 points lower.

GROUPS contains 28000 labeled faces, but a detailed analysis of its characteristics reveals some dataset difficulties: 1) Contrary to the other two datasets most individuals present a single image in the collection; 2) the mean inter-eye distance, 25 pixels, is respectively the half and the quarter than those present in LFW and MORPH faces; and 3) the subject pose and the illumination conditions are much less controlled.

### Table 1. Facial based GC state-of-the-art accuracies in large datasets.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Dataset</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[8]</td>
<td>GROUPS</td>
<td>89.8%</td>
</tr>
<tr>
<td>[12]</td>
<td>GROUPS</td>
<td>90.4%</td>
</tr>
<tr>
<td>[9]</td>
<td>GROUPS</td>
<td>90.8%</td>
</tr>
<tr>
<td>[38]</td>
<td>LFW</td>
<td>94.8%</td>
</tr>
<tr>
<td>[40]</td>
<td>LFW</td>
<td>98.0%</td>
</tr>
<tr>
<td>[15]</td>
<td>LFW</td>
<td>97.2%</td>
</tr>
<tr>
<td>[23]</td>
<td>LFW</td>
<td>96.9%</td>
</tr>
<tr>
<td>[35]</td>
<td>MORPH</td>
<td>97.1%</td>
</tr>
</tbody>
</table>

2.2. **Ocular GC**

Compared to the face, that has indeed large tradition among the facial analysis community, the ocular region has certainly received less attention in the GC problem. This is however not the case of identity recognition based on the periocular area, that aims at exploiting useful non-iris features of that area as suggested by [37] and [3]. In this sense, there are already interesting results with challenging datasets as reported by [28]. Compared to the face trait, there is a significant accuracy in identity recognition using remarkable less facial information, as argued by [4]. Different authors claim that ocular based systems are able to cope with challenging scenarios where the whole face is not available, e.g. partially occluded, or even when available, it has been altered with plastic surgery, as evidenced by [32, 29]; and [24].

Centering on the GC problem, iris may be an alternative to facial pattern, as pointed out by [41]. However, these systems require higher resolution images and more user cooperation, limiting the application scenarios.

There are however some results of periocular GC, that we briefly summarize here. The validity of the ocular area for GC was already suggested in video processing, as described in [11]. More recently similar conclusions have been achieved with image
datasets. In [29] an experiment was carried out with a collection of less than 1000 faces reaching a GC accuracy of 85% with local binary patterns (LBP) features.

A slightly larger dataset is considered in [27], where 4232 images of 404 subjects were analyzed in terms of gender and ethnicity, reaching for gender up to 97% of accuracy with local descriptors, comparable results to using the facial region. Certainly, the dataset contains high quality images (ocular area of $250 \times 250$ pixels).

A slightly harder scenario is tackled in [25] where the FERET dataset, see [33], is used with a periocular region of $74 \times 88$ pixels. However, the final dataset evaluated contains just 200 images reaching an accuracy around 90%. Certainly this dataset has already achieved very high GC rates with facial features (close to 100%), see [5] and [2].

Even if the reported results are promising, we conclude that the literature on periocular GC is limited, as the existing results were obtained with reduced datasets, and/or high resolution images. We therefore argue the interest of the analysis in the wild, where we compare the periocular GC performance with state-of-the-art facial GC systems on the currently hardest dataset, i.e. GROUPS.

3. System overview

The proposed system works with periocular images. This area is extracted after normalizing the face in terms of scale and rotation. Given the rough eye location annotations, the normalized facial image is obtained automatically after rotating, scaling and cropping the original images to force the ground truth eye locations to the specific eye positions $(16, 17)$ and $(42, 17)$, where the inter-eye distance is of 26 pixels, rather similar to the mean inter-eye distance in GROUPS, resulting an image of $59 \times 65$ pixels. The periocular region considered in the rest of this paper, has similar inter-eye distance but reduces the pattern dimension to $49 \times 19$ pixels. An example of the normalized face (F) and its corresponding periocular (P) pattern are illustrated in Figure 2.

![Fig. 2. Face (F) and corresponding periocular (P) patterns. Each face image is normalized so that the center of each eye is placed at a fixed location. The original face pattern has a resolution of $59 \times 65$ pixels with the eyes located in $(16, 17)$ and $(42, 17)$. The resulting periocular pattern has a resolution of $49 \times 19$ pixels with the eyes located in $(11, 11)$ and $(37, 11)$.](image_url)
As features, we focus on local descriptors, that are currently well known in the facial GC literature. Given an image, \( I \), each descriptor provides a feature vector, \( x \). This feature vector is obtained over a cell grid on the pattern, see an illustration in Figure 3. The use of the cell grid integrates spatial information in the feature vector that would be lost if the whole pattern is considered as a single region.

Summarizing, the process for each image, \( I \), is as follows: First, the grid resolution is established defining \( c_x \) horizontal and \( c_y \) vertical cells, extracting the features for the set of \( c_x \times c_y \) non overlapping cells. Cell \( i-th \) is described using a histogram, \( h_i \), where the bins contain the number of occurrences of the different descriptor codes present in the particular cell, following a Bag of Words scheme as defined by [14]. For a particular descriptor, \( D \), the final resulting image feature vector, \( x^D_I \), contains the concatenation of \( c_x \times c_y \) histograms, i.e. the feature vector is defined as \( x^D_I = \{h_1, h_2, ..., h_{c_x c_y}\} \). The feature extraction procedure is graphically outlined in Figure 3.

In the experiments, each descriptor performance is individually evaluated to determine the best grid configuration for the GC problem. As we do not know in advance the best grid resolution configuration, and observing the input pattern dimensions, each descriptor is evaluated for different grid resolutions in the periocular area, from \( 1 \times 1 \) up to \( 8 \times 6 \) cells. A known experimental framework is used, the protocol described in [15], for better comparison of the results with previous literature.

It is out of the scope of this paper to define a new classification paradigm, but to better design an automatic GC system with existing tools. For that reason, and taking into account our previous experience in face GC, we make use of Support Vector Machine (SVM) classifiers with RBF kernel, see [43].

The classifier, given a feature vector in a n-dimensional feature space, \( x \in \{x_1, x_2, ..., x_n\} \), assigns a label, \( y \), given the GC bi-class problem \( y \in \{\pm 1\} \), corresponding each label to the particular class in our problem: male or female.

Finally, once the optimal grid configuration for each single descriptor is found, further experiments are described in the experimental section, making use of score level fusion.
4. Features

As indicated above, most GC approaches focus on the facial pattern. Recent state-of-the-art facial GC approaches are mainly based on local descriptors. We study some of these local descriptors in the periocular scenario. More precisely, we have considered:

- Histogram of Oriented Gradients (HOG by [16]),
- Local Binary Patterns (LBP by [34]),
- Local Ternary Patterns (LTP by [39]),
- Weber Local Descriptor (WLD by [13]), and
- Local Oriented Statistics Information Booster (LOSIB by [19]).

They are briefly described in the following subsections.

4.1. Histogram of Oriented Gradients

HOG was introduced by [16] as a technique to represent the gradient orientations in a regular area of the image, called cell. As mentioned, an input image is divided into a rectangular grid of cells, representing each cell by a histogram, and the whole image by the concatenation of the respective cell histograms.

HOG reduces the illumination influence normalizing each cell histogram taking into account the cell neighborhood, that is known as the block.

In their work, Dalal and Triggs made use of a cell size of $8 \times 8$ pixels. As exposed in the previous section, we will consider different cell dimensions. In the experiments presented below, we make use of the implementation by [26] that considers a block of $2 \times 2$ cells, and 9 bin histograms.

4.2. Local Binary Patterns

LBP is a powerful texture descriptor, that has extensively been applied to real world Computer Vision problems, as exposed in [34]. A main LBP feature is its robustness to monotonic gray-scale changes.

In their original proposal, each pixel is encoded using its $3 \times 3$ neighborhood. This definition assigns a pixel value as a binary code after comparing its gray value with its $p$ neighbors within radius $r$, (the original definition uses $p = 8$ and $r = 1$), as follows:

$$LBP_{p,r} = \sum_{k=0}^{p-1} s(g_c - g_k)2^p$$

(1)

where, $g_c$ and $g_k$ are respectively the gray levels of the center pixel, and the $p$ neighbors, where $(k = 0, 1, \ldots, p - 1)$. The function $s(x)$ is defined as:
\[ s(x) = \begin{cases} 
1, & x \geq 0 \\
0, & x < 0 
\end{cases} \]

LBP is therefore computed easily. In texture analysis, the image is later described using a histogram. However, to process faces, and avoid the complete loss of features spatial information, since the work by [1], the normalized face is represented by the concatenation of histograms of a cell grid, similarly to HOG.

The LBP success has given rise to a large number of variants. Among them, we firstly mention uniform LBP, \( LBP^u \). This representation instead of considering the 256 different codes of the original proposal, uses only the subset of patterns that has been empirically demonstrated to be more frequent. A LBP code is uniform if the binary pattern contains a maximum of two transitions to bit-level, from 0 to 1 or vice versa. To illustrate this, patterns of 00000000 (0 transitions), number 00110000 to (2 transitions) and 11100111 (2 transitions) are uniform, while the patterns of 11001001 (4 transitions) and 01001010 (6 transitions) are not uniform. Among the 256 original codes, just 58 are uniform. They cover more than 90% of the code appearances. The inclusion of a codification for non uniform patterns, makes a total number of 59 codes, reducing almost 5 times the feature vector dimension.

4.3. Local Ternary Patterns

LTP, developed by [39], generalizes LBP getting a more discriminant and less sensitive to noise texture descriptor. This is done considering as threshold value instead of center pixel gray level used in LBP, a range around that value. This action reduces the sensitive to noise in almost uniform areas, quite presents in facial images, and smooths light illumination gradients. To do so, the LTP definition extends LBP considering three values for function \( s(x) \).

\[
s(x) = \begin{cases} 
1, & x > k \\
0, & k \geq x \geq -k \\
-1, & x < -k 
\end{cases} \]

where \( k \) is the constant threshold used to reduce the influence of noise. Typically instead of using 3\(^p\) codes the codification considers the coding in two halves, the positive and negative. Therefore if a ternary code were 1100-1-100, the upper and lower codes would respectively be 11000000 and 00001100.

4.4. Weber Local Descriptor

WLD is inspired by the Weber’s Law that states the size of a noticeable difference, is proportional to the original stimulus value, see [13]. This law observes that human perception depends on the stimulus change, but also on the original stimulus intensity. In this sense, the Weber fraction relates the increment threshold, \( \Delta I \), and the initial stimulus intensity, \( I \), i.e. \( k = \frac{\Delta I}{I} \).

The implementation considers two components: 1) the differential excitation, that relates the intensity differences of the pixel neighborhood and the center pixel, and the original image; and 2) the gradient orientation. Similarly to other descriptors a concatenated histogram is used for representation.
4.5. Local Oriented Statistics Information Booster

LOSIB, described recently by [19], is a descriptor enhancer. It is based on LBP, but the main difference is that it computes the local oriented statistical information in the whole image.

Given a $3 \times 3$ neighborhood, the intensity differences for the 8 neighbors/orientations may be computed as follows:

$$d_k(x_c, y_c) = |g_k - g_c|$$

being $k = 0, 1, ..., p - 1$. LOSIB computes the mean of all differences along the $p$ orientations for the $m \times n$ image pixels:

$$v_k = \frac{\sum_{x_c=1}^{m} \sum_{y_c=1}^{n} d_k(x_c, y_c)}{m \cdot n}$$

Describing the image by means of the $p$ mean values, i.e. $\{v_0, v_2, ..., v_{p-1}\}$. We have adapted the original definition for texture processing to face analysis concatenating the histograms obtained from a grid of cells.

5. Experimental evaluation

As mentioned above, the experimental evaluation follows the protocol defined by [15]. This protocol defines a 5-fold cross-validation setup, containing the subset of faces belonging to GROUPS that present an inter-eye distance larger than 20 pixels. This subset has a total number of 14385 facial images.

To illustrate the classification performance, we make use of the ratio between the correct predictions and all the predictions made, i.e. $\text{Accuracy} = \frac{TP}{TP+FP}$, where $TP$ and $FP$ stand respectively for the number of correct and incorrect predictions.

We remind the reader that the periocular images have been normalized to an inter-eye distance of 26 pixels, with a final dimension of $49 \times 19$ pixels, see Figures 2 and 3.

5.1. Single descriptors classification

In a first analysis, we present the evaluation of the mentioned local descriptors on the periocular region. This study considers six local descriptors: HOG, LBP, LBP$u^2$, LTP, WLD, and LOSIB. All of them are evaluated with 48 different grid resolutions, $cx \times cy$.

The parameters $cx$ and $cy$ stand respectively for the number of cells in the horizontal and vertical axis. They are analyzed in the range $[1, 8]$ and $[1, 6]$, i.e. analyzing grid from $1 \times 1$ to $8 \times 6$ cells, making a total of $6 \times 48 = 288$ variants.

Given a normalized image, as summarized above, for each descriptor configuration a feature histogram is composed by the concatenation of the respective cell histograms, computed the corresponding rectangular $cx \times cy$ grid. Its performance is evaluated using a RBF Support Vector Machine Classifier (SVM), with $C$ (trade-off between margin and error) and gamma parameters fixed respectively of $C = 1$ and $\text{gamma} = 0.07$. The results, computed exclusively for the first Dago’s fold, are presented in Table 2, clustering in color tones the different accuracy intervals.
Table 2. Periocular based single descriptor accuracy results (%) obtained for the first Dago's protocol fold. For each descriptor, the best grid configuration is highlighted, while the cell color serves to cluster accuracies.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>57.44</td>
<td>62.28</td>
<td>64.46</td>
<td>63.78</td>
<td>67.48</td>
<td>66.93</td>
<td>68.78</td>
<td>67.85</td>
</tr>
<tr>
<td>2</td>
<td>64.05</td>
<td>69.70</td>
<td>72.03</td>
<td>74.39</td>
<td>76.51</td>
<td>77.03</td>
<td>77.34</td>
<td>77.06</td>
</tr>
<tr>
<td>3</td>
<td>69.84</td>
<td>72.24</td>
<td>74.20</td>
<td>77.06</td>
<td>79.77</td>
<td>79.63</td>
<td>80.31</td>
<td>79.90</td>
</tr>
<tr>
<td>4</td>
<td>71.82</td>
<td>75.66</td>
<td>78.36</td>
<td>80.31</td>
<td>81.92</td>
<td>82.10</td>
<td>82.44</td>
<td>81.96</td>
</tr>
<tr>
<td>5</td>
<td>73.67</td>
<td>76.66</td>
<td>79.60</td>
<td>80.38</td>
<td>82.03</td>
<td>81.72</td>
<td>83.02</td>
<td>82.30</td>
</tr>
<tr>
<td>6</td>
<td>74.26</td>
<td>77.60</td>
<td>80.60</td>
<td>80.38</td>
<td>82.03</td>
<td>81.72</td>
<td>83.02</td>
<td>82.30</td>
</tr>
<tr>
<td>LBP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>63.75</td>
<td>67.41</td>
<td>70.97</td>
<td>71.45</td>
<td>72.24</td>
<td>72.08</td>
<td>74.70</td>
<td>74.56</td>
</tr>
<tr>
<td>2</td>
<td>71.65</td>
<td>74.63</td>
<td>75.79</td>
<td>77.99</td>
<td>77.33</td>
<td>77.54</td>
<td>78.09</td>
<td>78.33</td>
</tr>
<tr>
<td>3</td>
<td>73.71</td>
<td>76.87</td>
<td>78.34</td>
<td>79.02</td>
<td>79.64</td>
<td>79.14</td>
<td>80.50</td>
<td>79.90</td>
</tr>
<tr>
<td>4</td>
<td>74.86</td>
<td>74.66</td>
<td>75.80</td>
<td>74.19</td>
<td>73.24</td>
<td>72.51</td>
<td>72.26</td>
<td>72.09</td>
</tr>
<tr>
<td>5</td>
<td>69.46</td>
<td>72.92</td>
<td>73.88</td>
<td>73.22</td>
<td>73.33</td>
<td>74.22</td>
<td>74.90</td>
<td>74.02</td>
</tr>
<tr>
<td>6</td>
<td>72.85</td>
<td>75.83</td>
<td>77.51</td>
<td>77.10</td>
<td>76.72</td>
<td>77.06</td>
<td>76.82</td>
<td>76.45</td>
</tr>
<tr>
<td>LTP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>64.77</td>
<td>68.19</td>
<td>71.03</td>
<td>70.97</td>
<td>71.45</td>
<td>72.24</td>
<td>74.70</td>
<td>74.56</td>
</tr>
<tr>
<td>2</td>
<td>71.65</td>
<td>74.63</td>
<td>75.79</td>
<td>77.99</td>
<td>77.33</td>
<td>77.54</td>
<td>78.09</td>
<td>78.33</td>
</tr>
<tr>
<td>3</td>
<td>73.71</td>
<td>76.87</td>
<td>78.34</td>
<td>79.02</td>
<td>79.64</td>
<td>79.14</td>
<td>80.50</td>
<td>79.90</td>
</tr>
<tr>
<td>4</td>
<td>74.86</td>
<td>74.66</td>
<td>75.80</td>
<td>74.19</td>
<td>73.24</td>
<td>72.51</td>
<td>72.26</td>
<td>72.09</td>
</tr>
<tr>
<td>5</td>
<td>69.46</td>
<td>72.92</td>
<td>73.88</td>
<td>73.22</td>
<td>73.33</td>
<td>74.22</td>
<td>74.90</td>
<td>74.02</td>
</tr>
<tr>
<td>6</td>
<td>72.85</td>
<td>75.83</td>
<td>77.51</td>
<td>77.10</td>
<td>76.72</td>
<td>77.06</td>
<td>76.82</td>
<td>76.45</td>
</tr>
<tr>
<td>LOSIB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>57.44</td>
<td>59.53</td>
<td>68.82</td>
<td>63.30</td>
<td>63.75</td>
<td>64.46</td>
<td>65.11</td>
<td>65.53</td>
</tr>
<tr>
<td>2</td>
<td>61.62</td>
<td>65.46</td>
<td>69.98</td>
<td>70.11</td>
<td>71.69</td>
<td>72.51</td>
<td>74.08</td>
<td>74.19</td>
</tr>
<tr>
<td>3</td>
<td>66.72</td>
<td>68.95</td>
<td>71.21</td>
<td>73.09</td>
<td>74.70</td>
<td>75.15</td>
<td>74.73</td>
<td>76.17</td>
</tr>
<tr>
<td>4</td>
<td>65.29</td>
<td>67.96</td>
<td>72.17</td>
<td>74.15</td>
<td>74.56</td>
<td>75.01</td>
<td>75.56</td>
<td>75.35</td>
</tr>
<tr>
<td>5</td>
<td>66.18</td>
<td>68.25</td>
<td>71.07</td>
<td>73.67</td>
<td>73.98</td>
<td>74.84</td>
<td>74.97</td>
<td>75.38</td>
</tr>
<tr>
<td>6</td>
<td>67.10</td>
<td>70.14</td>
<td>73.06</td>
<td>74.22</td>
<td>74.36</td>
<td>76.04</td>
<td>76.45</td>
<td>75.59</td>
</tr>
</tbody>
</table>

A first observation of the results, suggests that each descriptor behaves differently according to the grid configuration. In terms of accuracy, both HOG and WLD reported the best performance. The best accuracies per descriptor, and their corresponding number of features and Matlab implementation processing time are detailed in Table 3. HOG is both the fastest and reached the best accuracy with a grid of 7 × 6 cells. The second best accuracy is reported by WLD, but it is also the second slowest requiring 183 milliseconds to provide a classification result. It is easily observed a relation between the number of features and processing cost. Faster processing is of course achieved reducing the grid dimension and choosing another implementation language.

This conclusion is also evidenced by the corresponding ROC curves, see Figure 4. Observe that the figure plots for each descriptor exclusively the curve for the best grid configuration in terms of accuracy according to Table 2.
Fig. 4. ROC curves using the Dago’s protocol for different single descriptors based on features extracted of the periocular area. Only the results achieved for the best grid configuration are presented.

Table 3. Periocular based best single descriptor results in terms of accuracy (%) obtained for the first fold of the Dago’s protocol. For each descriptor the best grid setup is indicated. The number of features and mean processing time in milliseconds for a Matlab implementation are included.

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Grid</th>
<th>Number of features</th>
<th>Processing time</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOG</td>
<td>7x6</td>
<td>378</td>
<td>12</td>
</tr>
<tr>
<td>LBP u2</td>
<td>6x3</td>
<td>1416</td>
<td>58</td>
</tr>
<tr>
<td>LBP</td>
<td>6x3</td>
<td>1526</td>
<td>66</td>
</tr>
<tr>
<td>LTP</td>
<td>3x2</td>
<td>9216</td>
<td>250</td>
</tr>
<tr>
<td>WLD</td>
<td>6x3</td>
<td>4608</td>
<td>183</td>
</tr>
<tr>
<td>LOSIB</td>
<td>7x6</td>
<td>336</td>
<td>15</td>
</tr>
</tbody>
</table>

The last two positions in the ranking are for LBP and LOSIB requiring respectively grids of 6x3 and 7x6 cells. Both reported more than 6 points lower accuracy than using HOG or WLD.

5.2. Score level fusion

Once that we have established the best grid configuration for each descriptor, highlighted in Tables 2 and 3, a fusion approach is considered. This is done to confirm whether or not the different descriptors are extracting complementary information from the periocular region for this task.

Among the different fusion alternatives at feature, score, or decision level, we have adopted the score level (SL) fusion approach. Score level fusion has evident benefits in terms of processing cost, as each first level classifier may be parallelized. For this problem, this effect is added to the proven accuracy improvement as claimed in [8] and the reduction of ambiguous cases occurrences in [10].

The final SL fusion setup is composed by two classification stages, as illustrated in Figure 5. The first stage combines from two to six, n, of the best classifiers per descriptor selected in the previous subsection. Each classifier provides a score that is passed to the second stage. This second stage makes use of the n first stage classifiers scores, \{s_1, ..., s_n\} that are fed to a SVM classifier with RBF kernel.
Fig. 5. Illustration of two stage classification fusion architecture, with $n$ classifiers in the first stage whose scores are fed into a second stage “meta” classifier.

![Diagram](image)

Fig. 6. ROC curves using the Dago’s protocol for the top 5 ranking fusion combinations based on the periocular area, see Table 4. The best single descriptor is presented as baseline (HOG).

After evaluating all possible combinations among the first stage classifiers, \( \binom{6}{2} + \binom{6}{3} + \binom{6}{4} + \binom{6}{5} + \binom{6}{6} = 57 \), we present in Table 4 the mean accuracy achieved in the 5-fold experiment defined in the Dago’s protocol. The table contains the accuracies for each best selected single descriptor classifier, and the top 5 of the fused approaches. For all of them the reported accuracy corresponds to the highest value achieved varying the cost and gamma parameters respectively within the intervals $C = [0.5, 5]$ and $\gamma = [0.04, 0.08]$.

Considering the single descriptor results, the best accuracy for a single descriptor is obtained for HOG, similarly to the results achieved in the previous subsection for the first Dago’s fold. The second best result is again obtained by WLD, slightly more than 0.5 points lower.
Table 4. Mean accuracy for the Dago’s protocol achieved for each single descriptor and the top-5 best fusion combinations.

<table>
<thead>
<tr>
<th>Descriptor(s)</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOG</td>
<td>81.61</td>
</tr>
<tr>
<td>LBP</td>
<td>79.60</td>
</tr>
<tr>
<td>LBP</td>
<td>79.04</td>
</tr>
<tr>
<td>WLD</td>
<td>81.03</td>
</tr>
<tr>
<td>LOSIB</td>
<td>76.09</td>
</tr>
<tr>
<td>LBP$^u_2$</td>
<td></td>
</tr>
<tr>
<td>LBP$^u_2$ + LBP + LTP + WLD (Ptop5)</td>
<td>82.67</td>
</tr>
<tr>
<td>HOG + LBP$^u_2$ + WLD (Ptop4)</td>
<td>82.70</td>
</tr>
<tr>
<td>LBP$^u_2$ + LTP + WLD (Ptop3)</td>
<td>82.73</td>
</tr>
<tr>
<td>HOG + LTP + WLD (Ptop2)</td>
<td>82.80</td>
</tr>
<tr>
<td>HOG + LBP$^u_2$ + LTP + WLD (Ptop1)</td>
<td><strong>82.91</strong></td>
</tr>
</tbody>
</table>

Observing the performances for the SL fused classifiers, the results indicate that all descriptors are not needed. Certainly the extracted information is overcomplete as some descriptors are redundant, but some descriptors seem to be providing complementary information, allowing a slight final overall improvement. Indeed, the top 5 combinations reported quite similar accuracies, within a range of 0.24 points. This evidence is confirmed by the corresponding ROC curves depicted in Figure 6. They are quite similar, but certainly better than the ROC curve obtained for the best single descriptor, i.e. HOG.

Attending to efficiency details, the $P_{top2}$ combination has the advantage of requiring the computation of just three descriptors for the first stage, but $P_{top4}$ does not make use of LTP features, i.e. the slowest to compute descriptor.

In any case, the accuracy is certainly worse than the one achieved by state-of-the-art facial GC systems on the same dataset, but it is however just less than roughly 8 points lower, compared with Table 1. We can conclude, that periocular GC is feasible and may be of interest for the problem, as some other recent works have pointed out. We have proved this fact in the likely most challenging public dataset for the problem.

Figures 7 and 8 illustrate respectively female and male classification errors obtained with $P_{top1}$ combination that includes four descriptors: HOG, LBP$^u_2$, LTP and WLD. For a human observer it is not particularly simple to label all of them without errors.

5.3. Multi-scale score level fusion

The results achieved in the previous section for GROUPS, confirm the validity of the periocular area to provide reliable GC performance in large and challenging scenarios. We move now one step forward, and carry out a final study to evaluate whether or not the specialized periocular classifier may provide additional or complementary information to a state-of-the-art face based GC.

Fusion of face and ocular regions is not a new idea, as already pointed out in the recognition problem by [24] and [31]. Here we go further combining multi-scale information as we fuse, facial (F), head and shoulders (HS) and periocular (P) information for GC, see Figure 9. We remind the reader the work by [2], that computed features at different resolutions from the facial pattern, combining shape, texture and intensity descriptors to improve the overall performance. We do not exploit the resolution cue strictly here, but extract features making use of different grid configurations in the pattern analyzed.
This approach is designed based on our previous proposals that combines facial and head and shoulders features for this task, see [8] and [9]. However, those systems focused on the complete facial and its local context, i.e. patterns F and HS. In the first work, three different descriptors were adopted in the first stage, and their respective score fused in the second stage of the SL architecture. That fusion mechanism reached a performance up to 89.8% for the Dago’s protocol. The selected descriptors and pattern combinations were:

- FHOG: HOG concatenated histogram of the facial pattern (59 × 65 pixels) using a grid of 8 × 8.
- FLBP: LBP^α2 concatenated histogram of the facial pattern (59 × 65 pixels) using a grid of 5 × 5.
- HSHOG: HOG concatenated histogram of the head and shoulders pattern (64 × 64 pixels) using a grid of 8 × 8.

In the following we evaluate the combination of those descriptors with the above mentioned 6 local descriptors applied on the periocular area, see Table 3. Thus, the final combination fuses up to 9 scores in the second stage. This approach analyzes the pattern of interest at different scales, configuring a grid setup for each particular descriptor and region of interest. We call this approach multi-scale GC.
Table 5. Mean accuracies (in brackets per class female/male) for the Dago’s protocol with multi-scale information score level fusion.

<table>
<thead>
<tr>
<th>Descriptor(s)</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>FHOG+FLBP$^{u2}$+HSHOG</td>
<td>90.49 (90.17/90.82)</td>
</tr>
<tr>
<td>FHOG+FLBP$^{u2}$+HSHOG+PHOG+PWLD+PLOSIB (FHSPtop3)</td>
<td>92.39</td>
</tr>
<tr>
<td>FHOG+FLBP$^{u2}$+HSHOG+PHOG+PLBP$^{u2}$+PWLD+PLOSIB (FHSPtop2)</td>
<td>92.41</td>
</tr>
<tr>
<td>FHOG+FLBP$^{u2}$+HSHOG+PHOG+PLBP$^{u2}$+PLOSIB (FHSPtop1)</td>
<td><strong>92.46</strong></td>
</tr>
</tbody>
</table>

Table 5 presents as baseline the accuracy for a state-of-the-art facial GC. The table also includes the best three results achieved fusing with features extracted from the periocular region, i.e. from 2 up to 6 descriptors.

The best overall accuracy obtained is almost 2 percentage points better (92.46 vs. 90.49) than the best previous GC performance in the literature. This effect is shown by all three reported classifiers fusion integrating HS, F and P features. This performance represents a significant improvement reducing the classification error in roughly 20%. This accuracy was achieved within the range the SVM cost and gamma parameters $C = \{0.5, 5\}$ and $\gamma = \{0.04, 0.08\}$.

This fact is also confirmed with the respective ROC curves, see Figure 10, that compared to the facial based GC, exhibit a remarkable improvement. These results evidence that the fusion of descriptors obtained of the F, HS and P patterns, improves GC performance.

The reader may also observe that the accuracy results per class are quite balanced, being slightly better for the male class in the approaches not using the periocular area, but this situation is reversed (in average) when periocular features are also included. The latter is a new fact compared to traditional results of facial based GC, see [30].

Figure 11 presents some examples of HS patterns belonging to the first Dago fold, that were incorrectly classified by the best facial based gender recognizer. The introduction of periocular features convinced the classifier to alter its decision. Some problems due to partial occlusions and the use of accessories are circumvented thanks to the more detailed ocular information.

Different male and female classification errors obtained with the multi-scale fusion are presented respectively in Figures 12 and 13. These classification errors were produced processing the first fold of the Dago’s protocol. The accuracy achieved for this particular fold was 94.76 (respectively 93.85 and 95.70 for female and male classes; the only fold with better accuracy for the latter). This test set contains 1479 female and 1442 male samples (total 2921), the number of respective errors was 91 and 62.

Fig. 9. Sample patterns used in the multi-scale fusion. From left to right, head and shoulders (HS) (64 × 64 pixels), face (F) (59 × 65 pixels) and periocular (P) (19 × 49 pixels) regions. Samples from GROUPS.
Fig. 10. ROC curves using the Dago’s protocol. Comparison of state-of-the-art classification based on the face and its local context, with multi-scale fusion combining, HS, F and P features.

Fig. 11. Selection of samples (first row female, second male samples) from the first fold that were erroneously classified using just features from the F and HS patterns, but the inclusion of periocular features changed the classifier mind.

(roughly 50% more in the female class). For each gender, the 18 depicted errors have been sorted according to their scores, i.e. the left and top most pattern shown in Figure 12 corresponds to the annotated female subject that received the largest male score, as so on. Observing the top 18 worst classified female patterns, the presence of glasses, the age variations, and the cluttered background contained in the HS pattern might certainly have affected to fool the classifier. Among the male errors, beards are not present among the ten first errors.
6. Conclusions

In this paper, we have designed a solution that integrates the periocular region, proving firstly that it is reliable for GC in large and uncontrolled scenarios. The achieved performance in the wild using descriptors fusion reported an accuracy 8 points lower compared to a state-of-the-art face based GC system. This is done at similar pattern resolution with a notorious reduction in the
area of analysis, that is useful when the whole facial pattern is not visible.

Secondly, a further combination with facial and local context features, reached up to 2 points larger accuracy than nowadays literature results, suggesting a complementary information introduced by the periocular area specialization to solve the problem. This improvement reduces the GC error in almost 20 percentage points considering the currently most challenging dataset for the problem. As drawback, there is certainly an increase in the processing cost and number of features, whose selection was not the focus of this paper, but must be addressed in the next future, as well as the evaluation of cross-database performance.

References


Acknowledgments

Work partially funded by the Institute of Intelligent Systems and Numerical Applications in Engineering and the Computer Science Department at ULPGC.