MEG: Texture operators for Multi-Expert Gender classification

Modesto Castrillón Santana[a], Maria De Marsico[b], Michele Nappi[c], Daniel Riccio[d]

[a]*SIANI, Univesidad de Las Palmas de Gran Canaria, Spain*
[b]*Sapienza University of Rome, Italy*
[c]*University of Salerno, Fisciano (SA), Italy*
[d]*University of Naples Federico II, Italy*

Abstract

In this paper we focus on gender classification from face images. Despite advances in equipment as well as methods, automatic face image processing for recognition or even just for the extraction of demographics, is still a challenging task in unrestricted scenarios. Our tests are aimed at carrying out an extensive comparison of a feature based approach with two score based ones. When directly using features, we first apply different operators to extract the corresponding feature vectors, and then stack such vectors. These are classified by a SVM-based approach. When using scores, the different operators are applied in a completely separate way, so that each of them produces the corresponding scores. Answers are then either fed to a SVM, or compared pairwise to exploit Likelihood Ratio. The testbeds used for experiments are EGA database, which presents a good balance with respect to demographic features of stored face images, and GROPUS, an increasingly popular benchmark for massive experiments. The obtained performances confirm that feature level fusion achieves an often better classification accuracy. However, it is computationally expensive. We contribute to the research on this topic in three ways: 1) we show that the proposed score level fusion approaches, though less demanding, can achieve results that are comparable to feature level fusion, or even slightly better given that we fuse a particular set of experts; the main advantage over the feature-based approach relying on chained vectors, is that it is not required to evaluate

a complex multi-feature distribution and the training process: thanks to the individual training of experts the overall process is more efficient and flexible, since experts can be easily added or discarded from the final architecture; 2) we evaluate the number of uncertain/ambiguous cases, i.e., those that might cause classification errors depending on the classification thresholds used, and show that with our score level fusion these significantly decreases; despite the final rate of correct classifications, this results in a more robust system; 3) we achieve very good results with operators that are not computationally expensive.
Keywords: Automatic gender classification, face images, multi-feature classification, feature level vs. score level fusion

---

1. Introduction

Demographic classification of people appearing in photos as well as in (real time) videos is attracting increasing interest in the scientific community, and especially among biometrics researchers. The different possible uses span a wide range. Among the commercially attractive ones, we find the possibility to improve marketing strategies and recommendation systems, that can be better tailored to the present user without explicit inquiry (for an example of the possible impact, see [1]). Possible further applications relate to Human-Computer Interaction and Ambient Intelligence, by allowing tailoring interfaces or ambient services for classes of users. For instance, recognizing an elder user might trigger automatically a suited visualization of elements on the screen, or appropriate events in the environment [2]. Finally, as demonstrated in literature, it is possible to increase the accuracy of biometric recognition in forensic and security-related applications. It is interesting to point out that demographics, and gender in particular, are always mentioned in the first place in almost all works dealing with the so called *soft biometric traits*. Those traits are defined as soft biometrics, since they are not able to univocally identify a subject, but rather a subclass of the population. However, they can support new recognition strategies [3], though they somehow remind of the earliest biometrics approaches

derived from the work by Bertillon in the 19th century, namely *Bertillonage*. As a matter of fact, in one of the earliest mentions of soft traits in biometric recognition, Wayman [4] proposes their use just for filtering a large biometric database. Limiting the number of entries to search in a database can greatly improve the speed of the response. However, errors in filtering can degrade the recognition performance. A different strategy to exploit soft biometrics is in [5]. In this case they are used to improve the accuracy of recognition in both verification (identity claim, 1:1 matching) and identification (no identity claim, 1:N matching) modes, by using them in combination with strong biometric traits (e.g., fingerprints in the cited work) to enforce the system response. Soft biometrics are exploited to improve face verification also in the very recent work presented in [6].

The work in [7] goes further, by presenting experiments showing at which extent the preliminary determination of those demographics can improve the accuracy of identity recognition carried out by strong traits, e.g., face. In this case, there is neither filtering (in the sense of applying the same system to a subset of the gallery), nor a-posteriori enforcement of the response. Rather, the idea is to train different systems on different traits/combination of traits, and to choose the right one to submit the incoming sample. The approach presented in the cited paper entails a human-in-the-loop approach, where an operator submits the biometric sample to the most suited system in a set trained beforehand on different combinations of demographic traits. An alternative is presented in [8], where no human intervention is requested. The common outcome is a significant increase of performance. In this paper, we focus on the gender classification (GC) problem from face images.

It is to notice that the influence of demographics on human appearance cannot be sharply identified, and yet is to be taken into account [9]. Moreover, GC in turn can be affected by age and ethnicity. The first issue has been investigated in [10], while the second one can be also related to the more general problem of the "other-race" effect. This denomination refers to the fact that humans are less proficient in discriminating demographics of people from other races,

if not helped by elements like clothes or hair. This is often explained by psychologists by the "contact hypothesis" [11]. This hypothesis suggests that the effect occurs as a result of a longer and wider experience with one's own- versus other-race faces, especially during childhood, when cognitive categories are acquired and consolidated. A somehow similar hypothesis may be formulated for computational approaches, whose performance in GC can be positively affected by a suitable, ethnicity-balanced training set, when a training phase is required. As a matter of fact, the work in [12] demonstrates that ethnicity-specific gender classifiers can improve the GC accuracy in a multiethnic environment. It is interesting to notice that, at a different level of detail/classification (gender vs. specific subject) this result is conceptually similar to the mentioned work in [7]. Despite the specific demographics under investigation, the benchmark dataset should be fairly balanced with respect to each factor [13], or specific classifiers should be trained on specific features and then combined [8]. This is the main motivation for choosing EGA (Ethnicity, Gender and Age) dataset [14] as one of the datasets used in our experiments, since this aspect is especially cared of. All EGA images are annotated with corresponding demographics information, that, in the present study, are used as the ground-truth for assessing demographic classification performance. It is to underline that images in the collection come from other popular face datasets, and are selected as to maintain at a minimum the distortions due to pose, illumination and expression (PIE) to better concentrate on demographics. As we will show in experimental results, this kind of controlled conditions seems to produce such a kind of homogeneity of images to also hinder GC. The 2015 NIST evaluation on GC [15] evidences the difference between GC with constrained or controlled, and unconstrained or *in the wild* datasets. A clear example of the second group is *The images of Groups* (GROUPS) [16], that we use for further experiments.

We propose to address the problem of automatic GC from face images by a multi-expert approach. We investigated the most appropriate choice and combination strategy from a set of local operators, each able to capture different aspects of images, to achieve accurate GC. As for combinations, we tested both

feature level and score level fusion. According to current literature, the former is expected to provide more accurate classification [17]. The reverse of the medal is the use of more computational resources and of a most demanding training process, since the feature vectors obtained are of larger size, namely the sum of sizes provided by the single experts, unless a further expensive step of feature selection/learning is performed. This also calls for more samples during training. On the contrary, when exploiting score-level fusion, experts can be trained individually, on smaller vectors, and this makes the training process both easier and parallelizable, since it does not require to evaluate a complex multi-feature distribution. Using the scores provided by the single experts as elements in a new feature vector, for a further training/classification step, the resulting size is equal to the number of experts. According to the quality of achieved results, we might accept this as a good compromise between accuracy and cost.

Before proceeding, it is worth underlining that we did not aim at demonstrating the performance of either new operators or new fusion strategies. Rather our contribution can be summarized in the following three points. 1) The achieved performance demonstrates that, when suitably applied, score fusion can provide results that are comparable to those obtained by feature fusion. 2) We further take into account the number of uncertain/ambiguous cases, and even when the accuracy is similar in percentage, this number significantly decreases, i.e., less situations arise that possibly require manual decision. This means that the obtained system is overall more efficient. 3) The satisfying results are achieved by the use of quite light/popular operators, and we consider this a further added value. We find worth to especially underline the novelty of point 2). At the best of our knowledge, no investigation in literature has taken into account in a thorough manner the effect of different (combined) classification operators on the number of ambiguous responses. Of course, this depends on both the operator(s) exploited, and on the classification thresholds. Given a similar accuracy, this characteristic can be used to further differentiate among classification performance and appreciate a possible higher robustness under this point of

view.

This work extends [18] in three respects. First, we add GROUPS as a new dataset for our experiments. The use of this new testbed of a significantly different size allowed us to assess possible influence on the performances of larger scale datasets, and to analyze and discuss some issues regarding scalability, robustness (in particular, performance degradation/stability across different data sources), and computational costs. Second, we test a larger number of local operators. Third, a deeper analysis of robustness to classification ambiguity (uncertain cases) is presented. The paper continues as follows. Section 2 summarizes some related work. Section 3 shortly describes the operation of the different local descriptors used. Section 4 illustrates score computation algorithms as well as fusion strategies, and introduces the problem of ambiguous classification. Section 5 shortly describes the datasets exploited for the experiments. We present the results of our experiments in Section 6. Section 7 closes the paper by drawing some conclusion and sketching future work.

2. Some related work

Face sex is particularly relevant for human interactions and for this reason the cognitive mechanisms driving the process of GC, when carried out by humans, has been often investigated. Some interesting studies are mentioned in [19]. In [20] the authors discuss how face GC is an extremely efficient cognitive process, that is acquired early during childhood, able to achieve almost 100% correct guesses for frontal unkown pictures. A subset of the experiments presented in [21] try to identify which are the cognitively critical zones exploited by humans for GC. Unfortunately, there is no clear evidence of where such zones are located in the face. The study carried out in [22] uses a pixelation filter to reduce frontal pictures (28,672 pixels) of male and female faces to 7168, 1792, 448 and 112 pixels, in order to measure the minimum information required for correct GC. The study tests the existence of sex differences in face gender processing, and concludes that male faces are categorised more efficiently than female faces,

and that in general subjects are more efficient in categorising same-sex faces. Due to the possible applications of automatic GC mentioned above, this also become a topic to investigate in biometrics and other fields. The final goal is to extract human rules to reproduce them in automatic GC by computer. However, this is still an open research problem for both psychologists and computer scientists.

Several physical and/or behavioral traits have been explored to tackle the problem of computer GC. A complete survey of GC methods is out of the scope of our work. The interested reader can refer to the work in [23]. The used biometrics include in particular speech [24] [25], the shape of the full body, when available [26], and the way of walking (gait) [27]. Performance achieved by combinations of traits have also been investigated, e.g., gait and face appearance [28]. The most used trait, especially in approaches based on Computer Vision, is without a doubt the face, alone (see the recent [29]) or in combination with obvious candidates like hair and clothes [30]. When exploiting the face appearance, the proposed approaches range from relatively more complex ones to those especially targeted at real time applications. For instance, the proposal in [31] uses dense Scale Invariant Feature Transform (SIFT) descriptors in combination with shape context. The computational burden of the former is decreased by using AdaBoost to select only the most relevant features. The latter is a scale and rotation invariant local descriptor that discretizes and indexes the distances and orientations between all of the n points on a shape, where n is a freely chosen parameter [32]. Another interesting face-based approach entails learning Local Binary Patterns (LBP) for face-based GC [33].

Focusing on face-based GC we find many of the the same variety of approaches that are exploited to address recognition of a specific subject. Of course, they are suitably adapted to address the coarser problem entailing only the two classes *Female* vs. *Male*, and to try to grasp the best features to distinguish between the two. Geometric-based or appearance-based methods are used, even if the latter are much more frequent in literature; in addition, also some hybrid proposal can be found.

In geometric-based methods, classification is based on distances between pairs of *fiducial points* [19], which are important points/landmarks that identify relevant elements of the face, such as the nose, mouth, and eyes. In some respect, this approach resembles *Bertillonage*, one of the earliest attempts to identify people by measures derived from physical traits. Studies using human subjects aim at establishing the importance of these distances in discriminating gender, however it is not well clear which are the critical zones that humans consider to achieve the best accuracy [21]. Moreover, once identified, the points must be accurately extracted and possibly aligned. An attempt to build a discriminator from a number of 2D as well as 3D measures is presented in [34]. Simple distances between key points in the pictures, ratios and angles formed between key points in the pictures, and 3D distances derived by combination of full-face and profile photographs are analyzed through discriminant function analysis to find the best discriminators.

In [19], the author selects 40 manually extracted points, which are chosen to minimize the amount of error in their extraction. These points are used to calculate 24 fiducial distances and two of them, namely the interpupillary distance and the distance between the eye midpoint and the philtral ridges midpoint, are used as normalizing factors of horizontal distances and vertical distances respectively. The study exploits discriminant analysis and demonstrates that only five of such normalized distances explain over 95% of the gender differences of "training" samples and predict the gender of 90% novel test faces with various facial expressions. In the first case, the test is carried out with 52 pictures from 26 males and 26 females of both the ARPA/ARL FERET database and pictures taken in the author's own laboratory. However, the set is not ethnicity balanced: the images consisted of 47 Caucasians and 5 Asians. Prediction is tested on a second set of 57 frontal pictures (26 female, 31 male faces) with various facial expressions.

Appearance-based methods either exploit pixel intensity values directly, or after some operation or transformation. These methods are sensitive to PIE variations. Gaussian RBF kernel gives the best performance in the work in [35],

where pixel intensity values are directly input to train the SVM. The work in [36] rather uses simple pixels comparisons to find features to exploit an Ada Boost approach. Haar-like features are used for real-time gender and ethnicity classification of videos in [37].

A number of works rely on the above mentioned LBP introduced in [38]. The work in [39] exploits LBP for multi-view GC. LBPs have been often used for GC in combination with other information, e.g., shape features [40]. Some authors argue that not all LBP bins may be equally relevant for GC, and the work in [41] uses AdaBoost to learn the most discriminative ones. LBP also inspired a number of variants, e.g., Local Gabor Binary Mapping Pattern used in [42]. In this approach, a face image is first transformed and represented as a series of Gabor magnitude pictures (GMP) by applying multi-scale and multi-orientation Gabor filters. Afterwards each GMP is encoded as a LGBP image by a Uniform Local Binary Pattern (ULBP) operator. Each obtained LGBP image is divided into non-overlapping rectangular regions, and spatial histograms are extracted from such regions. Combining regional histograms together to obtain a final LGBP feature vector would produce a vector of too high dimension, therefore suitable reduction techniques are investigated.

A further popular candiate for appearance-based GC is represented by SIFT. These are used for example in [43] as features in a Bayesian approach to estimate the posterior probability of a face trait at a specific time, conditional on features identified in previous frames of a video sequence. Temporal dependencies are represented by a Markov model. Classification requires determining the maximum a posteriori class at a given time.

The hybrid approach in [44] exploits Haar wavelets as appearance features and uses AdaBoost algorithm to select the stronger ones. Geometry features are regarded as apriori knowledge that help achieving a better classification. Active appearance model (AAM) is exploited to locate 83 landmarks. The method gets 3403 geometry features, from which the 10 most significant ones are picked, normalized and fused with the appearance features.

More complex methods have been proposed in literature, with even slightly

better performance. However, our aim is to investigate the combination of simple operators and to test the best way to fuse their responses, in order to achieve the best possible results. Having selected operators with a low computational complexity that can be used in parallel, we aim at obtaining results close to the state-of-the-art approaches with time performance suited for real time applications. Moreover, it is worth further underlining the contribution given by our proposal not only in terms of net performance, but also in terms of thorough investigation on the effect of different classifier (combinations) on the level of uncertainty/ambiguity of results.

3. Local Descriptors/Experts

We consider the following collection of local descriptors, that have already been applied in different scenarios of facial analysis: 1) Local Binary Patterns (LBP) [45]; 2) Local Gradient Patterns (LGP) [46]; 3) Local Ternary Patterns (LTP) [47]; 4) Local Derivative Patterns (LDP) [48]; 5) Weber Local Descriptor (WLD) [49]; 6) Local Phase Quantization (LPQ) [50]; 7) Histogram of Oriented Gradients (HOG) [51]; 8) Intensity based Local Binary Patterns (NILBP) [52] 9) Local Salient Patterns (LSP) [53]; and 10) Local Oriented Statistics Information Booster (LOSIB) [54].

Local Binary Patterns - LBP. LBP operator has been initially proposed for texture classification. Afterwards, the work by Ahonen et al. [55] introduced its use as descriptor for facial analysis. The original definition encodes the center of a 3×3 window comparing its value with each of the neighboring ones. Each neighborhood pixel is assigned a 1 if its value is greater than the central pixel, and a 0 otherwise. The final central pixel code is produced concatenating the 1s and 0s of the neighborhood into a binary number. Typically, a histogram is used to represent the image. Given the above procedure, it is clear that LBP coding is computed easily and quickly, and has proven its discrimination power in different real world texture classification problems, thanks to its robustness to monotonic gray-scale changes. Since its initial introduction, LBP definition

has been modified in a number of ways. For instance, it has been extended to arbitrary circular neighborhoods of radius R with P neighbors. In particular, in order to achieve higher robustness in facial processing, they are computed in a localized way. The normalized facial image is divided into a grid, and the histograms computed from the single grid cells are chained to obtain the final feature vector. As an alternative, a *feature image* F can be obtained, by substituting each pixel in the original image I by its LBP code.

Local Gradient Patterns - LGP. The LBP operator has attracted lots of attention and multiple variants have been proposed recently. A number of alternatives are based on different comparison criteria and on different characteristics in the pixel neighborhood. LGP operator makes use of the neighborhood gradient values of a given central pixel. The gradient is computed as the absolute value of intensity difference between the central pixel and each of its neighboring ones. Compared to LBP, gradient values substitute pixel values in the neighborhood of the central pixel, while their average substitutes the value of the central pixel as a threshold; given these differences, LGP encoding is performed similarly to LBP. Even in this case, a feature image F can be obtained from I, by substituting each pixel with its LGP code.

Local Ternary Patterns - LTP. LTP extends LBP to 3-valued codes. Gray levels within a intensity range of width $\pm t$ around the gray level g_c of the central pixel are quantized to 0, those above $g_c + t$ are quantized to $+1$ and those below $g_c - t$ to -1. Since t is a user defined threshold, LTP codes may be more resistant to noise, but no longer invariant to the kind of gray level transformations tolerated by LBP. LTP operator produces a ternary code instead of a binary one. However, several authors split each ternary pattern into its positive and negative parts (Upper Pattern and Lower Pattern respectively). For representation purposes, they may be used as two separate channels of descriptors, and exploited for classification by computing separate histograms and similarity metrics. The latter can be used individually or be combined. In the same way, two different feature images F_U and F_B can be obtained.

Local Derivative Patterns - LDP. LBP may be considered as the represen-

tation of first-order circular derivative pattern of images, i.e., a micro-pattern of the binary gradient directions. LDP increases the coded information detail representing a higher-order local pattern, by encoding directional pattern features based on local derivative variations. The n^{th} order LDP encodes the $(n-1)^{th}$ order local derivative direction variations. On the one side, LBP encodes the relationship between the central pixel and its neighbors. On the other side, LDP templates are more complex, extracting higher-order local information by encoding various distinctive spatial relationships contained in a given local region. Given an image I, the first-order derivatives are denoted as I'_α where α=0°, 45°, 90° and 135°. Given g_c a point in I, and g_p, $p = 0, \ldots, P-1$ its neighbors, the four first-order derivatives at g_c can be written as:

$$LDP^1(g_c) = \left\{ \begin{array}{l} I'_{0^\circ}(g_c) = I(g_c) - I(g_3), I'_{45^\circ}(g_c) = I(g_c) - I(g_2), \\ I'_{90^\circ}(g_c) = I(g_c) - I(g_1), I'_{135^\circ}(g_c) = I(g_c) - I(g_0) \end{array} \right\} \quad (1)$$

The second-order directional LDP, LDP^2_α, in direction α is defined as:

$$LDP^2_\alpha(g_c) = \left\{ \begin{array}{l} f(I'_\alpha(g_c), I'_\alpha(g_0)), f(I'_\alpha(g_c), I'_\alpha(g_1)), \ldots \\ \ldots, f(I'_\alpha(g_c), I'_\alpha(g_7) \end{array} \right\} \quad (2)$$

where $f(.,.)$ is a binary function which determines the type of local pattern transition, and encodes the co-occurrence of two derivative directions at different neighboring pixels:

$$f(I'_\alpha(g_c), I'_\alpha(g_p)) = \left\{ \begin{array}{l} 0, if I'_\alpha(g_c) \cdot I'_\alpha(g_p) > 0 \\ 1, if I'_\alpha(g_c) \cdot I'_\alpha(g_p) \leq 0 \end{array} \right. \quad (3)$$

Finally, the second order Local Derivative Pattern $LDP^2(I)$ is the concatenation of the codes according to each direction. Higher order derivatives are computed in a similar way. For more details, see [48]. Feature images corresponding to the different derivatives can be obtained as above.

Weber Local Descriptor - WLD. WLD also encodes differences of pixel intensity within a local neighborhood. However, it is inspired by Weber's Law stating that human perception of a pattern depends both on the change of a stimulus and also on its original intensity. Accordingly, WLD comprises two components: differential excitation and orientation. Even in this case, the two components are computed for each pixel x_c by considering it as the center of a neighborhood. The former one is a function of the sum over such neighborhood of the ratios between the intensity differences with each of the neighbors, and the value of x_c (for details see [49]). The orientation component is the gradient orientation of the same x_c, and computed as in [56]. For a given image, both components make up a concatenated WLD histogram. For further details see [49]. As for WLD, we use the code publicly available[1], that produces the pairs of values $< differentialexcitation, orientation >$ using 8 values (3 bits) for the former and 12 (4 bits) for the latter. We pack each pair in a 8 bit value to produce a single gray level value, that is used to produce the feature image F.

Local Phase Quantization - LPQ. The codes produced by LPQ are insensitive to centrally symmetric blur (e.g., due to motion, or out of focus). Similarly to previous descriptors, it is computed locally at every pixel location, then encoding the image as a histogram. The method is based on the blur invariance property of the Fourier phase spectrum. Therefore, the local phase information is extracted using the short-term Fourier transform (STFT) computed over a square neighborhood at each pixel position x. For details see [50]. Feature images are computed as for LBP.

Histogram of Oriented Gradients - HOG. After dividing the input image into a rectangular grid of cells, this descriptor computes a histogram of the gradient orientations in each cell, representing the whole image by the concatenation of the respective cell histograms. The influence of illumination

[1] http://www.cse.oulu.fi/CMV/Research/NewTextureDescriptors

is addressed by normalizing each cell histogram taking into account the cell neighborhood, known as block. In the experiments presented below, we make use of the implementation by [57] that considers blocks of 2×2 cells, and 9 bin histograms. As for WLD, the values for the angle (4 bits) and the magnitude (4 bits) of the gradient are packed in a single 8 bit code (gray level) to produce the feature image F.

Intensity based Local Binary Patterns - NILBP. NILBP is another LBP variant that tries to reduce the LBP oversimplification of local structure. To do this, NILBP computes the difference of each pixel in the neighborhood with the neighborhood mean, μ, instead of considering as reference the gray value of the central pixel. The feature images are computed as for LBP.

Local Salient Patterns - LSP. This recent LBP redefinition focuses on the location of the largest differences within the pixel neighborhood, to remove the noise influence. The coding considers the possible pairs of neighbor indexes $(p_{diffmax}, p_{diffmin})$ that provide respectively the maximum and the minimum difference with the central value. Therefore there are 57 distinguished values (the last one corresponds to equal differences for all neighbors). This descriptor has reported better rates in different facial analysis. We have included 5 different variants of it in the study below. In particular, LSP_0 refers to computing the difference of each pixel with respect to the central one of the neighborhood, as described, and taking the histogram of values assigned to the central pixel (or chain of histograms, if the image is divided into cells). LSP_1 refers to computing the difference of each pixel with respect to the following one in the neighborhood. The code to assign to the central pixel is computed as above, and also histograms are computed in the same way as LSP_0. LSP_2 refers to the same procedure, where the difference is substituted by computing, for each p_i in the neighborhood, the (circular) value $p_i + p_{i+2} - 2p_{i+1}$. LSP_{01} refers to chaining the results of LSP_0 and LSP_1, and finally LSP_{012} refers to chaining the results of all three LSP_n. The feature image F is obtained by summing the codes from the three LSP_n for each pixel.

Local Oriented Statistics Information Booster - LOSIB. This texture

booster is based on LBP. The main difference is that it computes the local oriented statistical information in the whole image. To do so, the intensity differences in the 3×3 neighborhood are computed as follows:

$$d_k(x_c, y_c) = |g_k - g_c| \tag{4}$$

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

$$\nu_k = \frac{\sum_{x_c=1}^{m} \sum_{y_c=1}^{n} d_k(x_c, y_c)}{m \cdot n} \tag{5}$$

The image is described in terms of p mean values, i.e. $\{\nu_0, \nu_2, ..., \nu_{p-1}\}$. As for LOSIB, it was not possible to obtain a feature images with significant appearance. It is worth underlining that feature images are treated as gray level images in all respects, since they capture specific trends in the original ones. The lack of apparent significance suggests that any processing on the resulting images would not lead to any usable result.

4. Score Computation and Fusion Strategies

4.1. Score Computation

Score computation by Likelihood Ratio (LR). The Likelihood Ratio (LR) is used to evaluate the membership of a sample to a specific class, after learning the class statistics. It has been introduced in biometrics to separate the class of genuine probes (those belonging to users enrolled in the system), from that of impostor probes (those belonging to unregistered users). The authors of [58] experimentally assess that, consistently with the Neyman-Pearson lemma, if, when False Acceptance Rate (FAR) is fixed at Ψ, we can find a constant η which maximizes Genuine Acceptance Rate (GAR), then the LR test represents the optimal test to assign the score vector X to either genuine or impostor class.

However, as in other LR applications, optimality is constrained by the precision of genuine and impostor score distributions estimates. Given a face image in input, we use here LR to produce a gender-discriminative (*Male/Female*) score; we then compare such score with a threshold properly fixed in advance, and following the result of the comparison the system decides if the input face belongs either to the class *Male* or to the class *Female*. As already underlined, a training phase is needed to estimate $f_{Male}(x)$ and $f_{Female}(x)$ distributions, and classification performance depends on the quality of such training.

All the experts in the system version that exploit LR for score generation execute the same operation pipeline, with the only difference of the local operator O that each of them uses to extract relevant features from I and transform it into a feature image F. For each pixel (x, y) in the image F the training phase learns the probability distributions Pr_{Male} and Pr_{Female}. We use a supervised learning procedure, where the gender male / female is known for each face image in the training set. In order to avoid training bias, training and testing sets have obviously no intersection. During matching, each pixel in the feature image F produces its own partial score $s(x, y)$ that contributes to the calculation of the final total score. The partial score is computed according to the learned distributions using the standard formula for LR:

$$s(x,y) = 2 \cdot \frac{log(f_{Female}(F(x,y)))}{log(f_{Male}(F(x,y)))} \tag{6}$$

The partial score produced by Eq. 6 generally gets a negative value if the pixel votes for the class *Male* and a positive value otherwise. The higher is the absolute value of the assigned partial score, the greater is the confidence that we can assign to the vote of that pixel for a class. There is an area of uncertainty in the interval around the 0, for which the partial score can be considered noise, rather than a really useful contribution for the calculation of the final score s. For this reason, we fix a threshold th_p for the partial score (here it has been experimentally set to $th = 1.3$) . The final score is calculated as:

$$s = \frac{1}{S} \sum \delta(s(x,y)) \cdot s(x,y) \tag{7}$$

where δ is the Dirac function returning 1 only if $|s(x,y)| \geq th_p$ and $S = \sum_{x,y} \delta(s(x,y))$. Similar considerations hold for the global score. A negative value represents a classification in the *Male* class, while a positive value represents a classification as *Female*. Even for the global score we can consider as uncertain/ambiguous those cases when the returned value is too close to the border between classes. Here, we deal with ambiguity of global scores returned by individual operators, and with ambiguity of fused scores too.

Score computation by Support Vector Machines (SVM). In [18], we evaluated using SVM classifier with either linear or RBF kernels. It is interesting to underline the contrasting results on the two datasets. As for EGA, since linear kernel was the one achieving the best performance for most operators, and this was confirmed with the added operators too, RBF is omitted from the results presented in the present work. As for GROUPS, the opposite trend is observed, therefore results by RBF are reported. The trade-off between margin and error, i.e., parameter C, was always fixed at $C = 1$. Even in this case we considered a decision threshold of 0 (negative vs. positive values). A SVM-based classifier (both based on a single operator or on a combination) outputs a score that also indicates the proximity of the sample to the threshold, and thus, as for LR-based classification, might be further used to evaluate the possible quality or ambiguity of the individual classification .

4.2. Feature Level and Score Level Fusion

There are different approaches to fuse the information provided by alternative experts. We may consider fusion either at feature, matching score, or decision level. On the one side, feature level (FL) fusion retains most information, but it is usually computationally more demanding due to the increase in the feature vector length. Moreover, the larger size also causes to require much more samples for methods based on a training phase. On the other side, decision level (DL) fusion looses too much information before the final result. Therefore, particularly when the number of experts to combine increases, score level (SL) fusion achieves the best compromise among, speed, preserved information and

performance.

In this work, we evaluate multi-expert systems using three different fusion protocols.

F-SVM performs feature level fusion. In particular, it exploits Support Vector Machines (SVM). A single linear SVM is trained on the feature vectors obtained by combining those produced by the single operators. In our case, they are obtained by stacking the histograms produced by the above described methods. More formally, given the set of experts, $\Omega = E_1, E_2, ..., E_n$, the protocol produces a new composed vector by combining the whole set of feature vectors $\Phi_{E_{i,k}} = f_{E_{1,k}}, f_{E_{2,k}}, ..., f_{E_{n,k}}$ extracted from a given image I_k by the individual experts E_i . Notice that each $f_{E_{i,k}}$ is a vector of variable size, depending on the corresponding operator. Therefore the size of the final vector grows according to the number of experts in a way that may hinder an effective and efficient classification.

S-SVM also exploits SVM but uses score level fusion. It entails using several first stage SVMs, each one trained on a different kind of feature vectors; even in this case the individual feature vectors are represented by histograms produced by the above methods. The protocol collects the responses of the individual experts E_i for a given image I_k, and then feeds them to a second stage SVM classifier. More formally, given the set of experts $\Omega = E_1, E_2, ..., E_n$, and their respective returned scores $s_{i,k}$ for image I_k, a new feature vector is composed as $\Sigma = s_{1,k}, s_{2,k}, ..., s_{n,k}$, and fed to a preliminarily trained linear SVM. Notice that each $s_{1,k}$ is a single score, therefore the size of the final vector is exactly equal to the number of experts.

S-LR uses score level fusion in conjunction with Likelihood ratio (LR). The single experts E_i produce their responses (scores) for a given image I_k using the feature images produced by the adopted operators. The individual scores are computed by LR, and afterwards the S-LR protocol combines them by examining them in pairs and selecting the best pair. More in detail, given a set of experts $\Omega = E_1, E_2, ..., E_n$, each of which produces a score $s_{i,k}$ for the feature image computed from I_k, for each possible pair $(E_{i,k}, E_{j,k})$ with $i \neq j$,

S-LR checks if both experts have voted for the same class (*Male*, *Female*), or $sign(s_{i,k}, s_{j,k})$. If this is true, the pair of experts is assigned a value of $s_{i,j,k} = sign(s_{i,k}) \cdot \sqrt{s_{i,k} \cdot s_{j,k}}$, which represents the fused score. Otherwise, the protocol assigns the value $s_{i,j,k} = 0$ to the pair. At the end, the protocol S-LR selects the pair of experts that provides the maximum $s_{i,j,k}$ in absolute value, or $s_{global,k} = Max_{i,j}(|s_{i,j,k}|)$.

4.3. The Role of Ambiguous Answers

All the classification protocols described so far provide a score as output. For a bi-class problem, the score sign can identify the class to which the sample was automatically assigned (*Male* or *Female*). In our case, males are associated to negative scores, and females to positive ones. Some samples obtain a score close the the border value, i.e. 0. This circumstance suggests a possibly high degree of uncertainty of the corresponding response. It is to point out and underline that this ambiguity is not an inherent characteristic of the classifier, but can affect from time to time single responses. In the same way, a very accurate recognition system might provide in some cases less trustworthy responses, due to temporary adverse conditions or to especially hard probes. Due to the singular nature of such phenomena, they usually remain uncovered when analyzing system performance, which are usually evaluated from aggregate statistics (e.g., True Positive or True Negative rates). However, it is very important to quantify the amount of this kind of responses too, to better appreciate the system quality. It is possible to define a threshold th_s that allows the system to indicate whether the response can be considered reliable, if $abs(s_{global,k}) \geq th_s$, or not, otherwise. In the second case we consider the response ambiguous. Those answers represent particularly complex cases for the system. In environments demanding high classification accuracy, they should be treated separately and differently. In the following we plot the curves relating the percentage of ambiguous answers and the accuracy (percentage of correct classifications) of the system versus variations of threshold th_s (see Figure 4 in Section 6.2 for EGA, and Figure 5 in Section 6.3 for GROUPS). It is worth underlining that strate-

gies producing a similar correct classification rate can differ by the number of ambiguous responses: of course, for similar accuracies, the lower this number, the better. In practice, by discarding such responses and avoiding considering them in computing system accuracy, the performance may further increase. In any case, the lower their number the more robust is the system behavior. As for the moment, we do not further process ambiguous responses. During real operation, the way they are treated depends on the kind of application. For instance, in a real time setting, they could trigger an alert to a possible human operator, who is in charge of either taking the final decision, or repeat the capture, or definitively discard the probe.

5. The Image datasets

Two different datasets are considered below for this particular problem, to provide conclusions in different scenarios of applications.

5.1. EGA

The aim of EGA (Ethnicity, Gender and Age face database) is to support experiments on face demographics. To this aim, it has been designed and implemented to provide demographics balance among dataset images, as well as flexibility even along time. EGA integrates into a single dataset face images extracted from different publicly available databases, in order to create a more heterogeneous and representative dataset. In most cases, even if a database is publicly available, it cannot be transferred. Therefore, in order to avoid copyright infringement, EGA has been conceived as a set of links and of annotations. Links connect to files previously processed by appropriate scripts, while annotations are provided to organize images according to demographic features such as ethnicity, gender and age. Each researcher can ask and obtain on her/his own the original datasets making up EGA. The scripts will reorganize and rename all requested images, according to the structure that was devised for EGA. In this way, it is possible to easily reconstruct the whole dataset, or parts of it,

but even to expand it, as new datasets become available and after they are annotated. In the present version of EGA, images are taken from CASIA-Face V5 [59], FEI [60], FERET [61], FRGC [62], JAFFE [63], and the Indian Face Database [64]. In particular:

- **CASIA-Face V5** includes images captured in a single session by an USB camera; image resolution is 640×480, 16 bit color. Faces are captured at different distances and can present illumination and pose variations, and eyeglasses; most subjects are young and of Eastern ethnicity;

- **FEI** includes 14 colour images per subject; images resolution is 640×480; faces have been acquired on a white background in FEI Laboratory in Brazil, and belong to subjects from 19 to 40 years old, mostly of Latin ethnicity;

- **FERET** dataset contains images of with resolution 256×384 for 8 bit greyscale; faces are categorized in sets (fa, fb, dup I, dup II) according to pose and acquisition period, and present slight variations in illumination and expression; the dataset is heterogeneous with respect to ethnicity, gender and age;

- **FRGC** includes images captured in controlled and non-controlled conditions; image resolution is high: 1704×2272 for 24 bit color; most subjects are of Caucasian ethnicity, and the number of subjects of other ethnicities is quite marginal; subjects are also mainly concentrated in a same age range (young/adult), while an adequate number of subjects is present for each gender;

- **JAFFE** contains images mainly gathered for facial expression analysis; image resolution 256×256 for 8 bit greyscale; subjects are all female and Japanese of apparently uniform age;

- **Indian Face Database** contains subjects in frontal pose with eleven different looking directions for each individual, plus some additional image

when available; the dataset is divided into two folders, one for male and one for female subjects; each image is 640×480 pixels, with 256 grey levels per pixel, and is captured on a uniform background, with four expression variations; all subjects are of Indian ethnicity, with an adequate distribution with respect to gender, but not with respect to age.

For sake of space, it is not possible to report here complete information about each dataset. The interested reader can refer to the original papers for further. However it is worth pointing out here that, being acquired at different times, with different equipment and in different settings, each dataset has its own characteristic specifications. This adds a further element of challenge to the final EGA collection. Just because it was collected with the aim to support experiments based on demographic traits, EGA does not include all images from the above databases, but subsets allowing an overall good balance in demographics composition and a lower influence of factors different from demographics (e.g., pose, illumination and expression - PIE). The present version of EGA includes 469 subjects from five ethnicities: a) African-American (53), b) Asian (111), c) Caucasian (162), d) Indian (75), e) Latinos (68). For each ethnicity, subjects are chosen to achieve the best possible balance between males and females. Gender subgroups are further divided into three age groups: a) young, b) adult and c) middle-aged, with adult being better represented due to the composition of the original datasets. Table 1 summarizes numeric details about composition characteristics. More on EGA can be found in [14].

5.2. GROUPS

The dataset *The images of Groups* (GROUPS) [16] is considered, in the recent literature on the problem, the most challenging scenario for evaluating solutions to biometric classification problems *in the wild* [15, 3]. This collection contains about $28,000$ annotated faces with large variations in terms of illumination, resolution and pose.

The dataset is a collection of images of groups of people from Flickr images (see Figure 1). The collection was built upon gathering the results of

Table 1: EGA composition

Ethnicity		***Gender***		***Age***	
Caucasian	162	Males	89	Young	25
				Adult	50
				Middle-Aged	14
		Females	73	Young	20
				Adult	33
				Middle-Aged	20
Asian	111	Males	54	Young	34
				Adult	14
				Middle-Aged	6
		Females	57	Young	33
				Adult	19
				Middle-Aged	5
Indian	75	Males	49	Young	3
				Adult	37
				Middle-Aged	9
		Females	26	Young	4
				Adult	15
				Middle-Aged	7
Latinos	68	Males	34	Young	7
				Adult	19
				Middle-Aged	8
		Females	34	Young	8
				Adult	16
				Middle-Aged	10
African American	53	Males	20	Young	3
				Adult	13
				Middle-Aged	4
		Females	33	Young	16
				Adult	11
				Middle-Aged	6

three searches: 1)"wedding+bride+groom+portrait"; 2)"group shot" or "group photo" or "group portrait"; 3)"family portrait". Undesirable images were removed through a standard set of negative query terms. A maximum of 100 images are returned for any given image capture day, and the search is repeated for 270 different days. The collection consists of 5,080 images containing 28,220 faces (14, 549 females and 13, 671 males), each labeled with age and gender. The percentage of faces that can be automatically detected is about 86%, and this is the subset usually exploited to face biometric algorithms. Many faces have low resolution (the median face has only 18.5 pixels between the eye centers, and 25% of the faces have under 12.5 pixels). Given the source of the images, there

Figure 1: Top row presents an original GROUPS sample (500 × 375 pixels). Bottom row presents the extracted faces at normalized resolution (59 × 65 pixels). Given the face eye locations, the image is rotated, re-scaled and cropped to place the center of each eye at locations (16,17) and (42,17).

is a great amount of variation in many aspects. People often present (dark) glasses, face occlusions, or unusual expressions. As expected, the variabilities in GROUPS images provoke a decrease in the accuracy achieved, that therefore hardly reaches 90%.

The most commonly used experimental setup is due to Dago et al. [65], This work defines a 5-fold cross-validation using only the subset of faces that can be automatically detected and present an inter-eye distance larger than 20 pixels, i.e. 7241 female and 7133 male samples. For methods adopting this protocol on GROUPS, the largest accuracy in GC achieved by exploiting only features extracted from the whole facial pattern is 88.59% [66]. However, the fusion with features specifically extracted from other regions, such as the periocular and mouth ones, have recently reported an overall accuracy improvement [67].

In this work, we focus exclusively on the whole facial pattern.

6. Experiments and results

Below we summarize the proposed strategies in terms of descriptors and fusion policies, and the results achieved for both datasets. For each of them, we first singularly analyze the whole collection of descriptors, to evaluate their respective best grid configurations and their relative performance. Then, we evaluate SL and FL fusion, and report the combinations providing the best performance.

All results are reported in terms of accuracy, defined as the number of correct classifications in relation to the total number of samples processed, $Acc = \frac{(TM+TF)}{M+F}$, where TM and TF refer respectively to the number of correct *Male* and *Female* classifications, while M and F indicate the number of total male and female samples in the test set.

6.1. Summary of the explored classification strategies

For sake of readers, this section summarizes the core elements of the classification strategies that we have compared. Sections 3, 4 have presented respectively the descriptors that we combine, two possible choices for score computation, and fusion at different levels. Table 2 summarizes the most relevant information. Furthermore, Figure 2 shows the flow of both training and testing phases using the presented fusion strategies. As highlighted by the experiments reported in [18], fusion at score level using LR on feature vectors extracted from feature images (S-LR) achieves slightly worse or comparable results than S-SVM, yet with lower computational demand. This trend is consistent with the new experiments presented in this paper, as detailed in the following.

As a final comment, it is worth pointing out that SL combination does not increase the processing cost in multi-core architectures, as the experts may be computed in parallel. We have to consider the slowest expert and furthermore the considerably shorter additional time for final fusion, so that a comparable processing time w.r.t. to single operators may be reached.

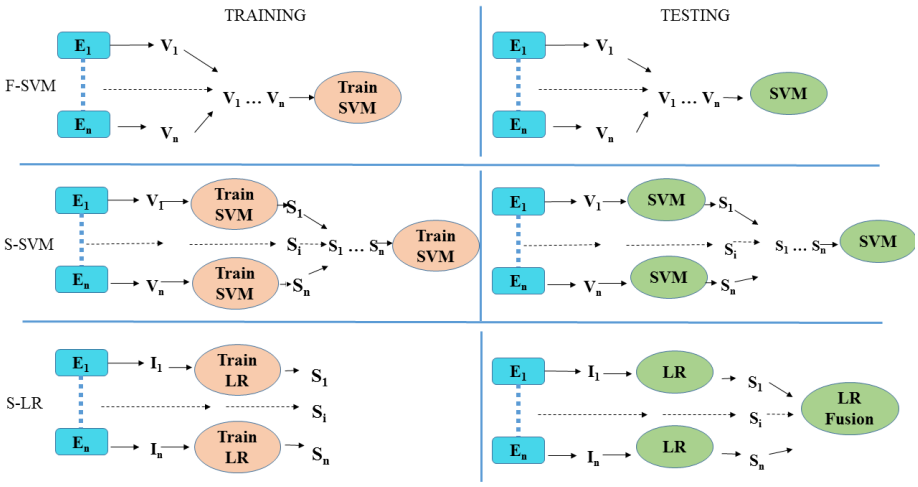

Figure 2: A schematic summary of the different fusion strategies. E= expert; V= feature vector; I= feature image; S= score

Regarding the classification times achieved by the single operators, we anticipate here that there is a time complexity factor which is intrinsic to the operator itself, and more specifically to the number of extracted features. Therefore consistent results can be observed for EGA and GROUPS datasets (see Table 4 and Table 7 below for numerical values). Differences related to generally higher classification times for EGA are due to the higher resolution of dataset images, so that EGA is more demanding for building the feature vectors.

6.2. EGA

For the experimental setup, a subset of EGA has been chosen with a single sample per identity with a random distribution between training and test sets. This is done to remove, in a relatively reduced dataset, the bias that might be produced by samples of the same identities in both sets. In summary 455 samples are analyzed in the evaluation protocol where 235 belong to the training set (respectively 111 female and 124 male samples), and 220 to the test set (respectively 103 female and 117 male).

After detecting the region of interest (ROI) containing the face using the

algorithm by Viola-Jones [68], each selected sample has been normalized to fix the centers of the eyes to specific locations, and rotated using such positions, with resulting normalized images of 64×100 pixels.

6.2.1. Single descriptor analysis

As mentioned above, for each expert we evaluated both linear and RBF kernels. However, we only report the values provided by the linear one for EGA, since it was in most cases the best, and those provided by RBF for GROUPS. The trade-off between margin and error, i.e., parameter C, was always fixed at $C = 1$. In all cases the accuracies are reported considering a decision threshold of 0 (negative vs. positive values).

We tested different configurations for the single operators, in particular the subdivision of the image according to different grid resolutions. Let us observe that the best grid dimension to use depends on the specific descriptor used. Due to the lack of space, detailed results concerning this aspect are included here only for the leading descriptor, namely LSP_{012}, that achieves the best overall performance on EGA. Table 3 illustrates the accuracy for each grid resolution ($rowheading \times columnheading$), with darker to lighter colors indicating values from worse to better.

It is possible to notice a diagonal trend in the dependence of accuracy on grid size. This indicates that there is no dominating relationship between width and height. There is a wide acceptable zone of possible combinations localized towards the upper left corner of the matrix, that shows an equivalent behavior of finer grids w.r.t. to coarser ones. The best performance of 96.36 is achieved by a grid with 7×5 resolution.

To summarize the results achieved using different configurations, we present in Table 4, for each descriptor, the best grid resolution, the number of features, the corresponding accuracies, and the average processing time per sample.

6.2.2. Combination of descriptors

As we expected, the fusion of more experts achieves better results. Table 5 presents the top-10 best combinations of subsets of descriptors. There is an evident improvement in relation to our previous work [18], reducing the classification error from 5.45% to 1.82%. This result is produced by the integration of a larger collection of descriptors, with the consequent possibility of a better choice.

In commenting the results, a first clear indication is that accuracy achieved by fusion at score level is generally better than that obtained by fusion at feature level, even if the best value is the same (98.18). It is also interesting to notice that the combinations of operators providing this result for F-SVM is a subset of those achieving the same best result for S-SVM, testifying the effectiveness of these combinations for both fusion choices. Regarding the different classification accuracies achieved on the two classes *Female* and *Male*, it is to point out that, though comparable, the former is systematically slightly lower that the latter, possibly indicating a slightly harder class due to higher variability (confusion).

It is worth exploring the results in more detail. In particular, the top row in Figure 3 presents the EGA images causing the classification errors produced by the best combination, i.e., the one that achieves the largest accuracy with the lowest number of descriptors, namely the set $\{WLD, LSP_2, LSP_{012}\}$. Before attempting a comment, it is worth reminding the reader that the training set contains samples achieving scores within the range ± 1. For the three leftmost test samples in the Figure, the classification returns a score close to the classification border (-0.10, -0.08, 0.01, respectively). This means that, having discarded such responses, the accuracy could have been even higher. However, the right most sample (score 0.78) achieves a score that is clearly far from being ambiguous. A detailed observation of the responses for the different experts points out that the expert community does not agree, but those experts considering the sample female got a higher value.

This example demonstrates that, considering that the SVM classifier pro-

Figure 3: Reminding that the classifier assigns the male label to those images with a score below 0: First row corresponds to EGA dataset classification errors obtained by the best fusion combination. From left to right, their respective scores are: -0.10, -0.08, 0.01 and 0.78. Second row presents a subset of the GROUPS classification errors obtained for the first fold: From left to right, their respective scores are: -1.54, -0.43, -0.10, 0.23 and 3.01.

duces a score as output, this value can also indicate the sample proximity to the decision threshold. As mentioned above, this value can be further used to evaluate the individual classification quality or ambiguity. The relation between accuracy and ambiguous responses is depicted for the best single descriptor, best SL and best FL fusion in Figure 4. In order to better interpret the figure, it is worth reminding the meaning and role of ambiguous samples. When identifying ambiguous samples is relevant, a further parameter is added to the classification system that defines a threshold th_s indicating whether the response can be considered reliable. The condition to meet is that $abs(s_{global,k}) \geq th_s$. Ambiguous answers represent complex cases possibly requiring special policies. The curves put in relation the percentage of ambiguous answers and the accuracy of the system versus variations of threshold th_s. Strategies producing a similar correct classification rate can differ by the number of ambiguous responses, and, of course, for similar accuracies, the lower this number, the better. For sake of simplicity, the plots avoid the explicit indication of the threshold. The increasing value of the threshold is implicitly reflected by the increased percent-

age of ambiguous answers. In other words, given the same set of images, the higher the threshold the higher the number of ambiguous answers; by discarding such responses, we experimentally demonstrate that the accuracy over the remaining ones increases, and this testifies the importance of identifying such cases. We plot the values only until the accuracy increases in a noticeable way. Furthermore, a system providing more than 70% ambiguous answers would be hardly useful. Notice that the accuracy values for a percentage of 0% ambiguous responses are those reported in Table 4 and Table 5.

The above results were obtained using chains of histograms as feature vectors. As for classification through feature images, we tested all the possible subsets of the set of considered operators, and fusion was always carried out at score level. In this case, grid subdivision is not applicable, since the feature image is treated as a usual gray level image. For sake of space, we do not report detailed performance analysis, since the obtained results were lower than those obtained using histograms, and therefore less interesting. In particular, the best result achieved by S-LR fusion is 93.30 and is obtained by the set $\{LBP, LDP, WLD, HoG\}$, while the best accuracy achieved by S-SVM fusion is 93.18 using the set $\{LDP, WLD, HoG\}$.

To compute the net improvement, ambiguous responses do not count in the denominator of the expression for Accuracy. It is evidenced that those responses that are farther from the classification border provide a better classification rate. In fact, the higher number of ambiguous responses is caused by an increased threshold, just meaning that the remaining answers are farther from the classification border. Therefore, we can point out the compromise achieved: a slightly lower number of useful responses vs. an increased classification precision. The plot evidences the additional positive effect not revealed by the accuracy numbers. Indeed the combination of experts increase accuracy while reducing the number of ambiguous cases.

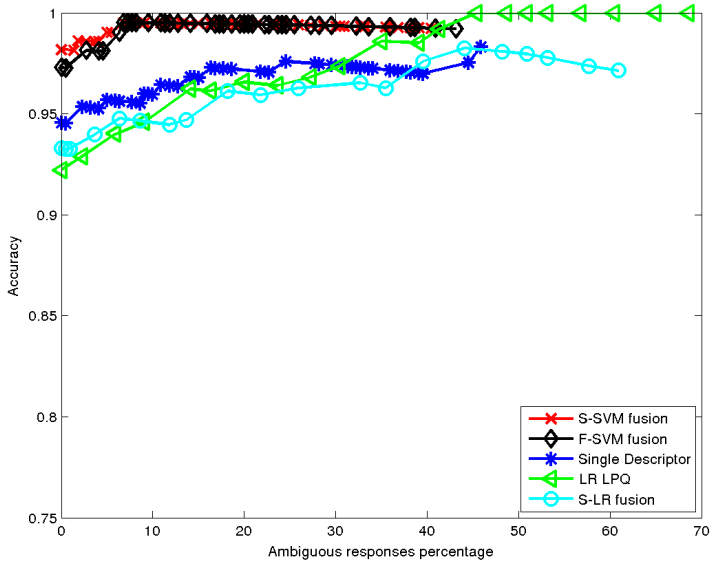

Figure 4: Ambiguous vs accuracy comparing the best results of single descriptors, of S-SVM and F-SVM fusion approaches, and of some LR-based approaches for the EGA dataset.

6.3. GROUPS

As mentioned above, we have used this dataset due to the fact that it is a popular dataset for assessing GC *in the wild*. Again the samples have been normalized according to the eye locations, which are included in the ground-truth information for the dataset, so obtaining normalized samples of 59×65 pixels.

We adopted the Dago's experimental protocol [65], that defines a 5-fold cross-validation that exploits the subset of larger faces in GROUPS, namely those that are automatically detected. Each fold contains respectively around 11500 training and 3000 test images.

6.3.1. Single descriptor analysis

Table 6 summarizes the accuracy for each grid resolution (*rowheading* $\times$ *columnheading*) for the best descriptor evaluating GROUPS. In the table, darker to lighter colors indicate values from worse to better accuracy.

It is interesting to comment the results in this table also making a comparison with Table 3. First, it is possible to notice a similar diagonal trend in the relationship between grid size and accuracy, but this trend is sharper and more evenly distributed over the matrix. In this case the region of best grid resolutions is more clearly identifiable, and the best result of 87.81 is achieved in the very corner of the matrix, i.e. 8×8. It is possible to conclude that, for both datasets, finer grids summarize less information yet preserving a better local description of details, and therefore produce a better accuracy.

Table 7 summarizes the best grid configurations, number of features and accuracy for single operators as obtained for the Dago's protocol applied to GROUPS. The grid configurations were chosen after analyzing the first fold. There is an evident performance decrease compared to EGA. The best accuracy achieved by a single descriptor is due to HOG, that reports a mean accuracy of 87.48%. The reader may observe the different behavior of the descriptors for a dataset of different nature.

6.3.2. Combination of descriptors

Once we have selected a grid configuration for each descriptor, we can evaluate the performance of their possible combinations. S-SVM fusion achieved better results with much lower computational effort, with an even more significant gap w.r.t. to F-SVM that the one highlighted for EGA in Section 6.2. For this reason, in order to provide the reader with a more clear and less cluttered set of results, Table 8 only report those ones. For S-SVM fusion, the best accuracy has been achieved by the combination of HOG, LGP and LPQ, reaching an accuracy of 89.22%. The same combination making use of F-SVM fusion reported a remarkable lower accuracy. This effect might be produced by the large number of features in the LGP feature vector, that seems to be affecting the benefits of the other two descriptors.

Regarding the difference in classification performance w.r.t. to *Female* and *Male* classes, compared to EGA the differences are often less marked and also in some cases inverted. This is reasonably due to the higher environmental

variations presented in images from GROUPS, that cause a similar increment in classification difficulty.

The above results were obtained using chains of histograms as feature vectors. As for classification through feature images, grid subdivision is not applicable. Similarly to EGA, even for GROUPS we tested all the possible subsets of the set of considered operators, with fusion at score level. Even in this case we only report the best obtained results. In particular, the best result achieved by S-LR fusion is 79.62 using $\{LBP, LTP_{low}, LPQ, LSP\}$, while the best accuracy achieved by S-SVM fusion is 79.73 using $\{LBP, LGP, WLD, LPQ, NILBP\}$. Given these results, which are significantly lower than those showed above, for sake of space we omit the full report.

It is interesting to notice an even more significant decrease of performance with respect to the dataset size, if compared with the use of histograms.

The variation of the percentage of ambiguous responses for the best single descriptor and for each fusion approach, and the way it influences classification accuracy, is illustrated in Figure 5. The plot interpretation is similar to Figure 4 for EGA. The best S-SVM fusion behavior is similar to the one exhibited for EGA. As a matter of fact, there is not just an increase in accuracy, but samples located farther from the class border areas get better classified. The plot confirms that the FL fusion of the operators involved in the best SL combination, decreases both accuracy and robustness against ambiguous responses.

Notwithstanding the similar trend, it is interesting to notice the more clear and direct relation between accuracy and ambiguity with respect to EGA. This is surely due to the more complex images in GROUPS, where higher variations in pose, illumination, and expression (PIE) increase the possibility to get a face image harder to classify.

6.4. Discussion

A first observation stems from the analysis of the curves resulting from the variation of two thresholds, one used for deciding the classification between the classes (*Male*/*Female*) and one set to detect possible ambiguity. The former is

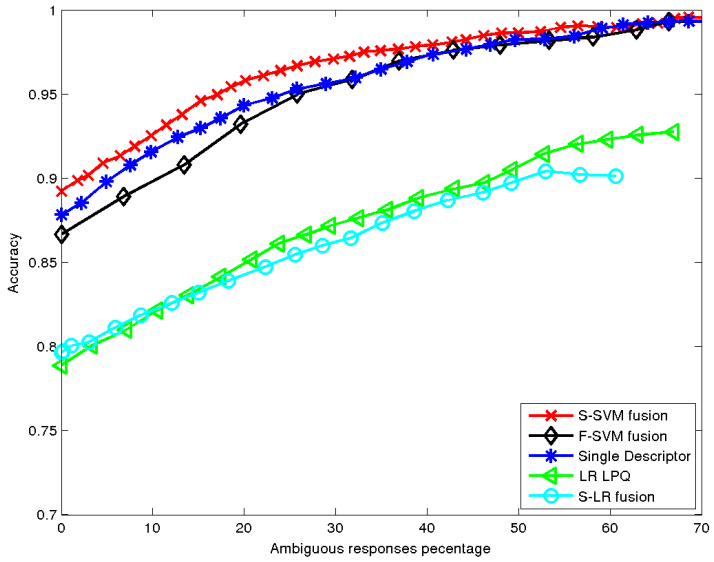

Figure 5: Ambiguous vs accuracy comparing the best results of single descriptors, of S-SVM and F-SVM fusion approaches, and of some LR-based approaches for the GROUPS dataset.

fixed to 0 for all the approaches as the distributions are normalized (negative vs. positive values). The second is used to discard the ambiguous responses. This action improves the accuracy of the final classification of remaining samples, but reduces the number of useful responses (accuracy is computed considering unambiguous responses only). As a second observation, we notice that given the fact that the accuracy is computed in relation to the number of useful responses, the lower the number of ambiguous responses the better the system, even when producing the same accuracy value.

The advantage of the fusion approaches is evident in Figures 4 and 5, which show that the multi-expert approach reduces the number of errors due to ambiguous cases. In order to maintain the figure readable, we only reported the curves corresponding to the best single classifier and best experts fusions. It is to point out the good performance of SL fusion. Due to the accuracy achieved, it would be preferred even with slightly worse accuracy with respect to FL,

since it is more flexible in terms of usability and cost, as it is better suited for parallelization.

Last but not least, when passing from a smaller dataset, with images captured in more controlled conditions, to a larger, *in the wild* one, we notice decreased performance yet similar trends and general behavior. While the former effect was expected, the latter one reinforces the validity of our results. In particular, we demonstrate that detecting and handling ambiguous responses can actually improve classification accuracy.

Results per class in both datasets indicate quite similar rates, but some combinations achieve a slightly worse performance for females.

Related to computational cost, EGA requires larger processing time to compute the features as the input images are larger. For EGA, WLD is clearly the most expensive descriptor: when fast processing is a demand, that descriptor should not be considered in any viable combination. For GROUPS, again WLD is the slowest descriptor to compute, but luckily enough the combinations achieving the best results do not use that descriptor. The system designed allows deciding which descriptors to combine to build a faster GC system.

7. Conclusions

In this paper we have analyzed a wide collection of local descriptors for the GC problem, increasing the variety of operators, and the nature of the evaluation datasets, with respect to our previous work. Our proposal deals with the fusion of the score output from the different operators, that we demonstrate offers a better trade-off between accuracy and cost with respect to fusion of features. We search for possible improvements in terms of both accuracy and reliability (number of ambiguous responses among the classification results). The experimental results have proven both benefits of combining more operators, i.e., the ability to increase accuracy while reducing the number of ambiguous situations. The latter allows the system to discard difficult samples while achieving a better relative performance. This effect is certainly achieved at the cost of reducing

the number of useful responses. However the fusion of more operators helps to also reduce the number of ambiguous/not valid samples. In this sense, even if similar accuracy is achieved in the overall dataset, a higher rate is obtained for the classified samples.

Acknowledgments

Work partially funded by the project TIN2015 64395-R from the Spanish Ministry of Economy and Competivity.

References

[1] D.-Y. Kim, X. Y. Lehto, A. M. Morrison, Gender differences in online travel information search: Implications for marketing communications on the internet, Tourism management 28 (2) (2007) 423–433.

[2] R. Casas, R. B. Marín, A. Robinet, A. R. Delgado, A. R. Yarza, J. Mcginn, R. Picking, V. Grout, User modelling in ambient intelligence for elderly and disabled people, Springer, 2008.

[3] M. Nixon, P. Correia, K. Nasrollahi, T. Moeslund, A. Hadid, M. Tistarelli, On soft biometrics, Pattern Recognition Letters 68, Part 2 (2015) 218–230.

[4] J. Wayman, A. K. Jain, D. Maltoni, D. Maio, Biometric Systems: Technology, Design and Performance Evaluation, Springer Verlag, 2005.

[5] A. K. Jain, S. C. Dass, K. Nandakumar., Soft biometric traits for personal recognition systems, in: ICBA2004, 2004.

[6] H. Zhang, J. R. Beveridge, B. A. Draper, P. J. Phillips, On the effectiveness of soft biometrics for increasing face verification rates, Computer Vision and Image Understanding 137 (2015) 50 – 62. doi:http://dx.doi.org/10.1016/j.cviu.2015.03.003.

[7] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, A. K. Jain, Face recognition performance: Role of demographic information, IEEE Trans. on Information Forensics and Security 7 (6) (2012) 1789–1801.

[8] M. De Marsico, M. Nappi, D. Riccio, H. Wechsler, Demographics versus biometric automatic interoperability, in: Image Analysis and Processing–ICIAP 2013, Springer, 2013, pp. 472–481.

[9] J. Bekios-Calfa, J. M. Buenaposada, L. Baumela, Revisiting linear discriminant techniques in gender recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (4) (2011) 858–864.

[10] G. Guo, C. R. Dyer, Y. Fu, T. S. Huang, Is gender recognition affected by age?, in: Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on, IEEE, 2009, pp. 2032–2039.

[11] P. Chiroro, T. Valentine, An investigation of the contact hypothesis of the own-race bias in face recognition, The Quarterly Journal of Experimental Psychology 48 (4) (1995) 879–894.

[12] W. Gao, H. Ai, Face gender classification on consumer images in a multiethnic environment, in: Advances in Biometrics, Springer, 2009, pp. 169–178.

[13] A. Torralba, A. A. Efros, Unbiased look at dataset bias, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011.

[14] D. Riccio, G. Tortora, M. De Marsico, H. Wechsler, EGA - Ethnicity, Gender and Age, a pre-annotated face database., in: 2012 IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS), 2012, pp. 1–8.

[15] M. Ngan, P. Grother, Face recognition vendor test (FRVT) performance of automated gender classification algorithms, Tech. Rep. NIST IR 8052, National Institute of Standars and Technology (April 2015).

[16] A. Gallagher, T. Chen, Understanding images of groups of people, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[17] A. Ross, A. K. Jain, Multimodal biometrics: An overview, in: Proc. 12th European Signal Processing Conference (EUSIPCO), 2004, pp. 1221–1224.

[18] M. Castrillón-Santana, M. D. Marsico, M. Nappi, D. Riccio, MEG: Multi-Expert Gender classification in a demographics-balanced dataset, in: 18th International Conference on Image Analysis and Processing (ICIAP), 2015.

[19] J.-M. Fellous, Gender discrimination and prediction on the basis of facial metric information, Vision research 37 (14) (1997) 1961–1973.

[20] V. Bruce, A. Young, In the eye of the beholder: the science of face perception., Oxford University Press, 1998.

[21] F. Gosselin, P. G. Schyns, Bubbles: a technique to reveal the use of information in recognition tasks, Vision Research 41 (17) (2001) 2261–2271.

[22] A. Cellerino, D. Borghetti, F. Sartucci, Sex differences in face gender recognition in humans, Brain research bulletin 63 (6) (2004) 443–449.

[23] C. B. Ng, Y. H. Tay, B.-M. Goi, Recognizing human gender in computer vision: A survey, in: Pacific Rim International Conference on Artificial Intelligence (PRICAI), 2012, pp. 335–346.

[24] K. Wu, D. G. Childers, Gender recognition from speech. part i: Coarse analysis, The journal of the Acoustical society of America 90 (4) (1991) 1828–1840.

[25] D. G. Childers, K. Wu, Gender recognition from speech. part ii: Fine analysis, The Journal of the Acoustical society of America 90 (4) (1991) 1841–1856.

[26] L. Cao, M. Dikmen, Y. Fu, T. S. Huang, Gender recognition from body, in: Proceedings of the 16th ACM international conference on Multimedia, 2008.

[27] X. Li, S. J. Maybank, S. Yan, D. Tao, D. Xu, Gait components and their application to gender recognition, Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 38 (2) (2008) 145–155.

[28] C. Shan, S. Gong, P. W. McOwan, Fusing gait and face cues for human gender recognition, Neurocomputing 71 (2008) 1931–1938.

[29] J. Bekios-Calfa, J. M. Buenaposada, L. Baumela, Robust gender recognition by exploiting facial attributes dependencies, Pattern Recognition Letters 36 (2014) 228–234.

[30] B. Li, X.-C. Lian, B.-L. Lu, Gender classification by combining clothing, hair and facial component classifiers, Neurocomputing 76 (1) (2012) 18–27.

[31] J.-G. Wang, J. Li, W.-Y. Yau, E. Sung, Boosting dense sift descriptors and shape contexts of face images for gender recognition, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, IEEE, 2010, pp. 96–102.

[32] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, Pattern Analysis and Machine Intelligence, IEEE Transactions on 24 (4) (2002) 509–522.

[33] C. Shan, Learning local binary patterns for gender classification on real-world face images, Pattern Recognition Letters 33 (2012) 431437.

[34] A. M. Burton, V. Bruce, N. Dench, et al., What's the difference between men and women? evidence from facial measurement, PERCEPTION-LONDON- 22 (1993) 153–153.

[35] B. Moghaddam, M.-H. Yang, Learning gender with support faces, IEEE Trans. on Pattern Analysis and Machine Intelligence 24 (5) (2002) 707–711.

[36] S. Baluja, H. A. Rowley, Boosting sex identification performance, International Journal of computer vision 71 (1) (2007) 111–119.

[37] G. Shakhnarovich, P. Viola, B. Moghaddam, et al., A unified learning framework for real time face detection and classification, in: Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on, IEEE, 2002, pp. 14–21.

[38] T. Ojala, M. Pietikäinen, T. Mäenpää, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. on Pattern Analysis and Machine Intelligence 24 (7) (2002) 971–987.

[39] H.-C. Lian, B.-L. Lu, Multi-view gender classification using local binary patterns and support vector machines, in: Advances in Neural Networks - ISNN, 2006, pp. 202–209.

[40] L. A. Alexandre, Gender recognition: A multiscale decision fusion approach, Pattern Recognition Letters 31 (11) (2010) 1422–1427.

[41] C. Shan, Learning local binary patterns for gender classification on real-world face images, Pattern Recognition Letters 33 (2012) 431–437.

[42] B. Xia, H. Sun, B.-L. Lu, Multi-view gender classification based on local gabor binary mapping pattern and support vector machines, in: International Joint Conference on Neural Networks (IJCNN), 2008, pp. 3388–3395.

[43] M. Demirkus, M. Toews, J. J. Clark, T. Arbel, Gender classification from unconstrained video sequences, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, IEEE, 2010, pp. 55–62.

[44] Z. Xu, L. Lu, P. Shi, A hybrid approach to gender classification from face images, in: Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, IEEE, 2008, pp. 1–4.

[45] M. Pietikäinen, A. Hadid, G. Zhao, T. Ahonen, Computer Vision Using Local Binary Patterns., Springer, 2011.

[46] B. Jun, D. Kim, Robust face detection using local gradient patterns and evidence accumulation., Pattern Recognition 45 (9) (2012) 3304–3316.

[47] X. Tan, B. Triggs, Enhanced local texture feature sets for face recognition under difficult lighting conditions, in: S. Zhou, W. Zhao, X. Tang, S. Gong (Eds.), Analysis and Modeling of Faces and Gestures, LNCS 4778, Springer Berlin Heidelberg, 2007, pp. 168–182.

[48] B. Zhang, Y. Gao, S. Zhao, J. Liu, Local derivative pattern versus local binary pattern: Face recognition with high-order local pattern descriptor, Image Processing, IEEE Trans. on 19 (2) (2010) 533–544. doi:10.1109/TIP.2009.2035882.

[49] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikainen, X. Chen, W. Gao, Wld: A robust local image descriptor, Pattern Analysis and Machine Intelligence, IEEE Transactions on 32 (9) (2010) 1705–1720. doi:10.1109/TPAMI.2009.155.

[50] V. Ojansivu, J. Heikkil, Blur insensitive texture classification using local phase quantization, in: A. Elmoataz, O. Lezoray, F. Nouboud, D. Mammass (Eds.), Image and Signal Processing, LNCS 5099, Springer, 2008, pp. 236–243.

[51] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: C. Schmid, S. Soatto, C. Tomasi (Eds.), International Conference on Computer Vision & Pattern Recognition (CVPR), Vol. 2, 2005, pp. 886–893.

[52] L. Liu, P. Fieguth, L. Zhao, Y. Long, G. Kuang, Extended local binary patterns for texture classification, Image and Vision Computing 30 (2) (30) 86–99.

[53] Z. Chai, Z. Sun, T. Tan, H. Mendez-Vazquez, Local salient patterns - a novel local descriptor for face recognition, in: International Conference on Biometrics (ICB), 2013.

[54] O. García-Olalla, E. Alegre, L. Fernández-Roble, V. González-Castro, Local oriented statistics information booster (LOSIB) for texture classification, in: International Conference on Pattern Recognition (ICPR), 2014.

[55] T. Ahonen, A. Hadid, M. Pietikäinen, Face description with local binary patterns: Application to face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (12) (2006) 2037–204.

[56] D. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Comnputer Vision 60 (2) (2004) 91–110, http://www.springerlink.com/content/h4l02691327px768/.

[57] O. Ludwig, D. Delgado, V. Goncalves, U. Nunes, Trainable classifier-fusion schemes: An application to pedestrian detection, in: 12th International IEEE Conference on Intelligent Transportation Systems (ITSC), 2009, pp. 1–6.

[58] B. Ulery, A. Hicklin, C. Watson, W. Fellner, P. Hallinan, Studies of biometric fusion., Technical Report IR 7346, NIST (2006).

[59] CASIA-FaceV5, http://biometrics.idealtest.org/.

[60] The FEI face database, http://www.fei.edu.br/ cet/facedatabase.html.

[61] P. J. Phillips, H. Wechsler, J. Huang, P. J. Rauss, The FERET database and evaluation procedure for facerecognition algorithms, Image and Vision Computing 16 (5) (1998) 295–306.

[62] P. Phillips, P. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, W. Worek, Overview of the Face Recognition Grand Challenge., in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2005.

[63] M. J. Lyons, S. Akamatsu, M. Kamachi, , J. Gyoba, Coding facial expressions with gabor wavelets, in: Proceeding of the IEEE International Conference on Automatic Face and Gesture Recognition, 1998, pp. 200–205.

[64] V. Jain, A. Mukherjee, The indian face database (2002).

[65] P. Dago-Casas, D. González-Jiménez, L. Long-Yu, J. L. Alba-Castro, Single- and cross- database benchmarks for gender classification under unconstrained settings, in: Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies, 2011.

[66] M. Castrillón-Santana, J. Lorenzo-Navarro, E. Ramón-Balmaseda, Descriptors and regions of interest fusion for gender classification in the wild. Comparison and combination with convolutional neural networks, ArXiv e-prints.
URL http://arxiv.org/abs/1507.06838v2

[67] M. Castrillón-Santana, J. Lorenzo-Navarro, E. Ramón-Balmaseda, Fusion of holistic and part based features for gender classification in the wild, in: New Trends in Image Analysis and Processing–ICIAP 2015 Workshops, Springer International Publishing, 2015, pp. 43–50.

[68] P. Viola, M. J. Jones, Rapid object detection using a boosted cascade of simple features, in: Computer Vision and Pattern Recognition, 2001, pp. 511–518.

Table 2: The core components used in the tested classification strategies with corresponding acronyms used in the text and a brief explanation

	Full denomination	**Acronym**	**Brief description**
DESCRIPTORS	Local Binary Patterns	LBP	Each pixel is used as a binarization threshold for its neighborhood and tis assigned the result string of binary vales; the result is a new image with the modified pixels, or an histogram of the obtained values; in the latter case a higher robustness can be obtained by chaining the histograms from a grid division.
	Local Gradient Patterns	LGP	Compared to LBP, gradient values substitute pixel values in the neighborhood of the central pixel, while their average substitutes the value of the central pixel as a threshold.
	Local Ternary Patterns	LTP – LTP_{high} – LTP_{low}	LTP operator produces a ternary code instead of a binary one; each ternary pattern is split into its positive and negative parts (Upper Pattern or *high*, and Lower Pattern or *low*, respectively).
	Local Derivative Patterns	LDP	While LBP encodes the relationship between the central pixel and its neighbors, LDP templates extract higher-order local information by encoding various distinctive spatial relationships contained in a given local region.
	Weber Local Descriptor	WLD	WLD comprises two components: differential excitation and orientation, that encode the change of stimulus with respect to its original intensity; even in this case, the two components are computed for each pixel by considering it as the center of a neighborhood.
	Local Phase Quantization	LPQ	The local phase information is extracted using the short-term Fourier transform (STFT) computed over a square neighborhood at each pixel position x.
	Histogram of Oriented Gradients	HOG	This descriptor divides the image into cells, computes a histogram of the gradient orientations in each cell, representing the whole image by the concatenation of the respective cell histograms.
	Intensity based Local Binary Patterns	NILBP	NILBP computes the difference of each pixel in the neighborhood with the neighborhood mean, instead of considering as reference the gray value of the central pixel.
	Local Salient Patterns	LSP	The coding considers the possible pairs of neighbor indexes ($p_{diffmax}$; $p_{diffmin}$) that provide the maximum and the minimum difference with the central value respectively; there are 57 distinguished values (the last one corresponds to equal differences for all neighbors).
	Local Oriented Statistics Information Booster	LOSIB	The main difference with LBP is that it computes the local oriented statistical information in the whole image.
SCORE COMPUTATION	Score computation by Likelihood Ratio	LR	All the experts execute the same operation pipeline, with the only difference of the local operator O that each of them uses to extract relevant features from I and transform it into a feature image F; for each pixel (x; y) in the image F the supervised training phase learns the probability distributions Pr_{male} and Pr_{female}; during matching, each pixel in the feature image F produces its own partial score s(x; y) that contributes to the calculation of the final total score; the partial score is computed according to the learned distributions using the standard formula for LR.
	Score computation by Support Vector Machines	SVM	A linear kernel is used; the SVM-based classifier can be either based on a single operator or on a combination.
FUSION	Feature Level Fusion by SVM	F-SVM	A single linear SVM is trained on the feature vectors obtained by combining those produced by the single operators; in our case, they are obtained by stacking the histograms produced by the descriptors.
	Score Level Fusion by SVM	S-SVM	It entails using several first stage SVMs, each one trained on a different kind of feature vectors (histograms produced by the descriptors); the protocol collects the responses of the individual experts for a given image, and then feeds them to a second stage SVM linear classifier that has been preliminarily trained.
	Score Level Fusion by LR	S-LR	The single experts produce their responses (scores) for a given image using the feature images produced by the descriptors; the individual scores are computed by LR, and afterwards the S-LR protocol combines them by examining them in pairs and selecting the best pair.

Table 3: Results obtained on EGA dataset for different grid configurations by LSP_{012} using SVM with linear kernel. This descriptor achieves the best accuracy using a single operator.

	1	2	3	4	5	6	7	8
1	74.09	74.55	81.82	85.00	80.45	87.27	84.55	86.36
2	75.00	75.91	82.73	84.09	86.36	90.45	90.00	88.64
3	76.36	83.64	87.73	89.09	91.82	95.00	90.91	92.73
4	82.73	86.82	88.18	92.73	94.09	95.91	91.36	93.18
5	81.36	85.91	90.91	90.45	94.09	90.91	90.91	92.27
6	79.09	87.27	86.36	90.00	92.27	94.55	94.09	92.27
7	79.09	87.27	88.18	91.82	**96.36**	94.55	92.73	94.09
8	83.18	86.82	90.45	90.45	94.09	93.64	91.36	94.09

Table 4: Best accuracy achieved per descriptor using SVM with linear kernel on EGA, mean processing time per image (milliseconds for a Matlab implementation in a i7 quad core processor with 4GB).

Feat.	HOG	LBP^{u2}	LBP	LTP	LGP	LPQ	WLD	LOSIB
Grid	8x8	7x8	7x7	4x6	4x6	7x4	8x8	5x6
# features	576	3304	12544	12288	6144	7168	16384	240
Acc.	91.82	93.18	91.82	94.09	90.91	94.09	95.91	90.00
t	22	59	168	224	216	128	1471	10
Feat.	NILBP	LSP_0	LSP_1	LSP_2	LSP_{01}	LSP_{012}	LTP_{high}	LTP_{low}
Grid	8x7	8x5	6x6	7x6	8x5	7x5	4x7	4x6
#features	3304	2280	2052	2394	4560	5985	7168	6144
Acc.	92.73	92.73	91.82	89.55	94.55	**96.36**	95.00	94.09
t	126	403	300	291	334	393	77	39

Table 5: Results on EGA for the best combinations of subsets of the considered operators using SVM with linear kernel and either SL or FL fusion (in brackets per class *Female*/*Male*).

Features combined	S-SVM Accuracy	F-SVM Accuracy
WLD LSP_2 LSP_{012}	**98.18** (98.06/98.29)	97.27 (97.09/97.44)
WLD LSP_1 LSP_2 LSP_{012}	**98.18** (98.06/98.29)	**98.18** (98.06/98.29)
WLD LOSIB LSP_2 LSP_{012}	**98.18** (97.08/99.15)	97.27 (97.09/97.44)
WLD LOSIB LSP_1 LSP_2 LSP_{012}	**98.18** (98.06/98.29)	**98.18** (98.06/98.29)
LGP WLD LSP_1 LSP_2 LSP_{012}	**98.18** (98.06/98.29)	**98.18** (98.06/98.29)
LGP WLD LOSIB LSP_1 LSP_2 LSP_{012}	**98.18** (98.06/98.29)	97.73 (97.09/98.29)
HOG WLD LSP_1 LSP_2 LSP_{012}	**98.18** (98.06/98.29)	97.27 (96.12/98.29)
HOG WLD LOSIB LSP_1 LSP_2 LSP_{012}	**98.18** (98.06/98.29)	97.72 (97.09/98.29)
WLD LSP_2	97.73 (98.09/98.29)	95.91 (95.12/95.73)
WLD LOSIB LSP_2	97.73 (96.12/99.15)	96.82 (96.12/97.44)

Table 6: Results obtained on GROUPS dataset (first fold) for different grid configurations by HOG using SVM with RBF kernel. This descriptor achieves the best accuracy using a single operator.

	1	2	3	4	5	6	7	8
1	61.62	63.41	70.87	71.82	72.82	75.28	75.25	76.86
2	68.98	72.44	74.91	75.76	76.82	78.19	78.70	80.04
3	70.18	74.22	77.82	79.60	81.07	81.72	81.89	83.19
4	71.14	76.28	79.56	81.34	82.54	83.26	84.35	86.41
5	72.00	77.34	80.49	82.13	83.43	84.15	84.94	86.41
6	73.40	78.57	81.31	83.36	84.53	85.04	85.52	86.88
7	73.88	78.91	82.88	84.29	85.45	85.93	86.20	87.50
8	74.25	80.38	84.01	85.35	86.07	86.37	87.02	**87.81**

Table 7: Best accuracy achieved per descriptor using SVM with RBF kernel on GROUPS, mean processing time per image (milliseconds for a Matlab implementation in a i7 quad core processor with 16GB).

Feat.	HOG	LBP^{u2}	LBP	LTP	LGP	LPQ	WLD	LOSIB
Grid	8x8	6x5	3x4	2x3	7x8	1x4	5x5	8x7
# features	576	1770	3072	3072	14336	1024	6400	448
Acc.	**87.48**	86.96	82.51	84.15	85.79	83.64	86.61	83.74
t	9	71	96	124	254	21	356	7
Feat.	NILBP	LSP_0	LSP_1	LSP_2	LSP_{01}	LSP_{012}	LTP_{high}	LTP_{low}
Grid	7x6	6x5	6x7	5x8	6x4	3x4	3x4	2x4
Acc.	87.13	85.66	86.58	83.43	86.61	86.16	85.79	85.48
# features	2478	1710	2394	2280	2736	2052	3072	2048
t	84	107	117	111	149	122	103	69

Table 8: Results on GROUPS for the best combinations of subsets of the considered operators using SVM with RBF kernel and SL fusion (in brackets per class *Female/Male*).

Features combined	S-SVM Accuracy
HOG LGP LPQ	**89.22** (89.22/89.21)
HOG LPQ NILBP	89.07 (89.24/89.11)
HOG LPQ LSP0	88.99 (88.84/89.16)
WLD NILBP LSP_{012}	88.92 (88.15/89.59)
HOG NILBP LSP_{012}	88.88 (88.53/89.15)
WLD NILBP LSP_1	88.80 (87.66/89.84)
LBP^{u2} WLD LSP_{012}	88.78 (89.22/89.51)
HOG LBP^{u2} LPQ	88.77 (88.64/88.90)
HOG NILBP LTP_ll	88.75 (88.28/88.99)
HOG LGP LSP_{012}	88.73 (88.13/89.34)