Università degli Studi di Salerno

Dottorato di Ricerca in Informatica e Ingegneria dell'Informazione
Ciclo 30 – a.a. 2016/2017

TESI DI DOTTORATO / PH.D. THESIS

Real-time face analysis for gender recognition on video sequences

ANTONIO GRECO

SUPERVISOR: **PROF. MARIO VENTO**

PHD PROGRAM DIRECTOR: **PROF. PASQUALE CHIACCHIO**

Dipartimento di Ingegneria dell'Informazione ed Elettrica
e Matematica Applicata
Dipartimento di Informatica

Contents

Fatti non foste a viver come bruti
ma per seguir virtute e canoscenza

You were not born to live like brutes
but to follow virtue and knowledge

- Dante Alighieri, Inferno, Canto XXVI -

Il successo è l'abilità di passare da un fallimento
all'altro senza perdere l'entusiasmo

Success consists of going from failure
to failure without loss of enthusiasm

- Winston Churchill -

Chapter 1

Introduction

1.1 Overview

The face is one of the most important parts of the human body, since it has some distinctive physical and expressive features which allow the identification of certain properties. For instance, by just looking at faces, humans recognize the gender and the ethnicity [1, 2], estimate the age [3], deduce the emotions and the state of mind [4, 5], determine if the person has a familiar face or is a stranger [6], and verify or recognize the identity of the individual [7, 8, 9]. Although all faces consist of the same parts in a specific spatial arrangement (the relative positions of the nose, eyes and others), primates have enviable abilities to use the subtle features and draw conclusions from faces in a remarkable and seemingly effortless operation. As a matter of fact, there is a neurophysiological evidence that the visual cortices of primates have single neurons that are selective to faces [10]. This fact demonstrates the importance that evolution gave to faces.

In recent years the variety and appeal of faces motivated several researchers to work on the problem of automatic face analysis coming from images. However, not a lot of studies have been produced to deal with the additional problems related to faces extracted in real life scenarios, when people are not aware of the presence of the camera. Although most of the attention has been devoted to still

images and not to videos, gender recognition is anyway considered among the most challenging problems [11].

1.2 Common challenges on still images

Although gender recognition from face images may appear a simple task, it is important to note that even human beings may find it challenging in certain situations. In fact, the study [12] demonstrates that the performance of humans in such a task reaches an accuracy lower than 95%.
In addition to the inherent difficulties of the topic, the automatic detection and analysis of a face is further affected by different problems. First of all, existing face detection algorithms achieve reasonable performance on frontal faces, but the accuracy gradually decreases when the face is tilted horizontally or vertically with respect to the camera.
Certain combinations of facial features also affect gender classification, as shown in Figure 1.1. The most challenging aspects are surely related to the pose variations and the partial occlusions of the face, for example with scarves, hats and glasses. While the former can be solved to some extent by normalizing the pose of the face using alignment algorithms [13], the latter is harder to solve due to the variety of all the possible occlusions. Furthermore, it has been demonstrated that the performance of gender recognition algorithms is strongly affected by the age of the people, as well as by their race or expression [14]. For instance, the wrinkles formed on elderly women may make their faces similar to elderly men. Figure 1.1 shows also other challenges, such as variations in the illumination and contrast.
In the last years, a great deal of literature has been produced with methods attempting to solve gender recognition from still images [11]. While significant progress has been observed, automatic systems have not yet reached the generalization capability needed to achieve good performance even in presence of variations in age, race, pose, illumination, and so on.

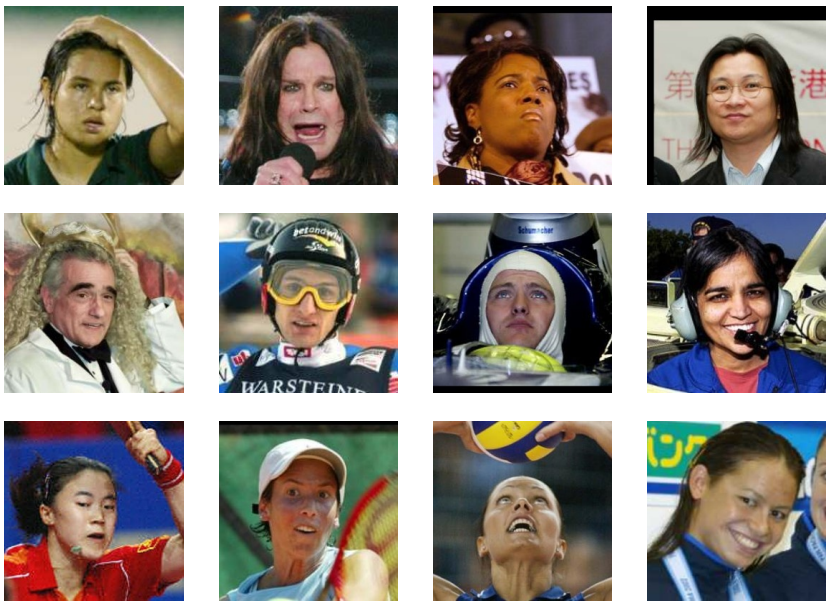


Figure 1.1 Typical problems of gender recognition systems: (first row) different expressions, ages or races; (second row) occlusions with wig, ski mask, balaclava, and microphone; (third row) different poses and illumination conditions.

1.3 Additional challenges in real life scenarios

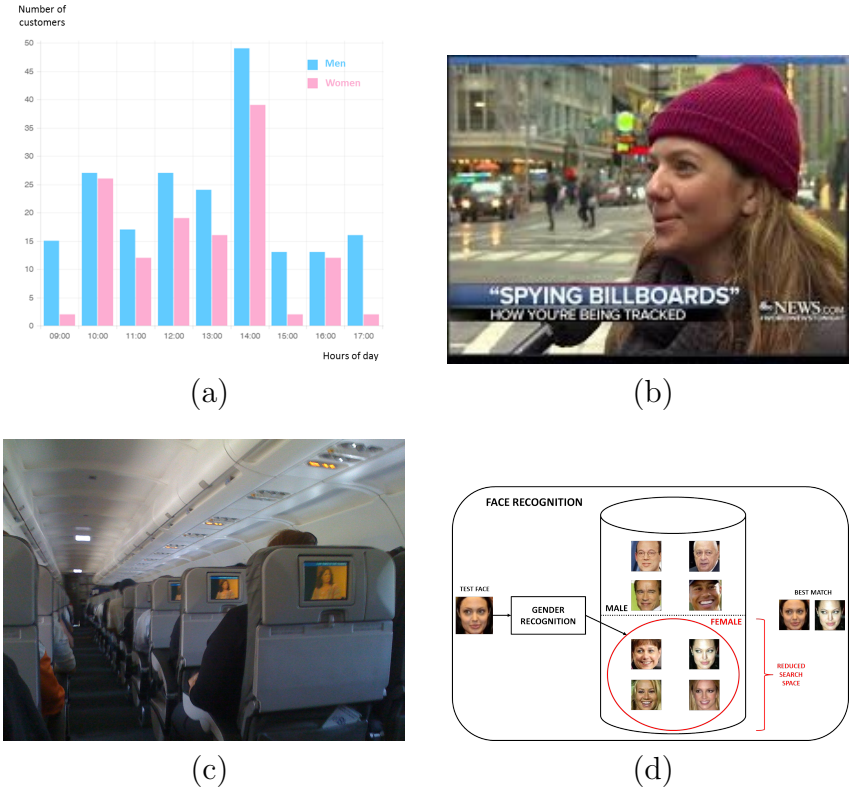
The algorithms for the automatic classification of gender have a lot of potential for commercial and security applications. Indeed, information such as gender, age and race are desired features that managers are eager to have for sophisticated market analysis. That information helps them to acquire more insights about the needs of customers with respect to the possible products that they can offer.

Typical examples in the retail field are given by the smart billboards or user interfaces which are able to modify their visual display depending on the gender of the person interacting with them. Another application could be tailored promotional material on screens installed in front of the passengers in taxis, trains

and planes. Moreover, retailers are nowadays considering to use face analysis algorithms to improve the shopping experience using cameras and customized web radios. Such smart systems will have the capability to personalize the promotional messages transmitted by the radio after recognizing the gender and the age group of the customers (eg. makeup advertising for women, toys for children). Figure 1.2(a-c) shows examples of some retail applications.

Another field where gender recognition algorithms can play an important role is video-surveillance. Indeed, there is a great demand for applications that are able to perform face recognition of suspicious individuals by analyzing images captured by surveillance cameras. The main challenge of these systems is the processing time needed for searching a match between the input face image and the thousands of samples stored in a reference database. One way for reducing the search space is to first detect the gender [15] and possibly the age group [16] and/or ethnicity [17] of the given face and then compare it only with the images which share the same properties in the database. Figure 1.2(d) depicts a high-level architecture of such a system.

When moving to these kind of applications, the images are acquired using traditional surveillance cameras and the algorithm has to process them in real time [18]. It implies that the problem of recognizing the gender of the persons becomes definitively more challenging.
Indeed, the algorithm has to deal with the fact that a person is not static but instead enters the scene, moves toward the camera and finally exits the scene. In other words, (1) the faces have to be found in real time, frame by frame, (2) even in presence of motion blur due to sudden movements of the persons or (3) in cases of pose variations. Indeed, the person is not collaborative, in the sense that he/she may be not aware of the presence of the camera and thus may not look towards the camera. (4) Furthermore, the movement of the person towards the camera implies that his/her face has a strongly variable dimension. Finally, (5) for each face, a single information concerning his/her gender has to be generated,



(a) (b)

(c) (d)

Figure 1.2 Examples of gender recognition applications. (a) A chart that shows the number of the daily visitors in a store, used to carry out market analysis (from *www.aitech.vision*). (b) A woman who talks about smart billboards. (c) Screens installed in front of the passengers in a plane, which can be equipped with smart cameras so as to show tailored promotional material. (d) The architecture of a system which performs a pre-classification of the gender in order to reduce the search space in large face databases.

thus the decision needs to be taken by discovering and then evaluating all the faces associated to a single person along the time and not just a single face.

1.4 The new trend of embedded vision

Although the above mentioned real-time processing constraint is probably the most challenging requirement of the algorithms which have to deal with image sequences acquired in real life scenarios, the current literature in this field does not dedicate to this problem the attention it would deserve. Indeed, in most of the cases the effort of the researchers is devoted to find the algorithm which maximize the gender recognition accuracy without considering the computational load required by the proposed method. The lack of experiments in this direction results in the compelling need of using very powerful processors to reduce the elaboration time of real applications which require a response in a few milliseconds.
The use of architectures with "unlimited" computing resources is in contrast with the growing interest towards general purposes embedded systems, able to strongly reduce the hardware cost of these kinds of solutions [19]. Indeed, the usage of an embedded system, especially if integrated directly on board of the cameras, strongly reduces the bandwidth required for transferring the video stream to an external server, as well as the energy consumption of the whole system, since a powerful server is no more required.
Such modern architecture, named embedded vision, together with the lower cost of the embedded system if compared with traditional server, strongly reduces the costs of the hardware infrastructure, making this kind of application more attractive for the retailers. Furthermore, the embedded vision implies a high scalability of the whole architecture: if the number of required cameras increases, it is not necessary to buy a new expensive server but just a couple of low cost embedded devices.
It is important to highlight that the design of an algorithm suited for embedded systems is a very challenging task. Indeed, all the efforts aimed at achieving a higher accuracy require also a greater computational burden. For example, higher resolutions of the images, and then of the single faces, which obviously preserve more details about the facial landmarks, require a higher processing time. Or else, a high number of images per person, needed for in-

creasing the reliability of the whole system as well as for extracting additional information, such as the persistence time of a person in front of the camera, requires a high frame rate.
So the research of a gender recognition solution suitable for embedded systems leads to the absolutely not trivial problem of finding out the best possible trade-off between the computational cost (and the resources required for the elaboration, both in terms of time and space) and the accuracy (the percentage of correct classifications) of the algorithm.

1.5 The lack of standard datasets

While in the last years several datasets have been proposed for benchmarking gender recognition algorithms, to the best of our knowledge there are no datasets which have become standards de facto. Also, there is not a standard way for evaluating algorithms on the existing datasets, as in most of the cases the partitioning of training and test sets is not clearly defined.
The difficulty is mainly due to the fact that the face datasets are generally collected for face recognition or verification. In recent years FERET [20], LFW [21], Adience [18], MORPH [22], Youtube Faces Database [23], PIE [24], Multi-PIE [25], AR [26], PUT [27], FEI [28] have been used for several face analysis purposes. However, only for LFW and Adience the labels for gender and the way for evaluating the algorithms are provided, but in most of the scientific papers the protocol is not respected, so making hard the comparison of the performance of a new algorithm with the existing ones.
All the abovementioned datasets consist of still images (or face images extracted from videos, as in the case of Youtube Faces Database) captured with professional cameras or, anyway, in situations where the person is aware of the presence of the camera. Such condition is very different from the video-surveillance scenarios, where the quality of the images is very poor due to the use of traditional surveillance cameras and to the wide range of vari-

ations in the appearance of the face image of an unaware person. ChokePoint [29], SCFace [30] and FR_SURV_VID [31] are composed of face images acquired in video-surveillance environments and have been used in the last years for face recognition purposes. Nevertheless, no labels or experimentation protocols are provided for gender recognition.

1.6 Contribution

This research has the purpose of investigating the possibility to perform gender recognition on face images extracted from video sequences acquired with surveillance cameras in real scenarios. As discussed above, these constraints impose to deal with additional challenges, such as real-time processing, high accuracy even in presence of motion blur and pose or scale variations, and the design of a mechanism to track the person in the scene, in order to provide a single classification of the gender, instead of one for each face image.
As explained in Section 2.1 the convolutional neural networks (CNNs) are able to achieve significant performance in terms of gender recognition accuracy, but they require computational and memory resources that are not available on low cost embedded devices. Indeed, the most effective CNN for face analysis, namely VGG-Face, requires more than 10 seconds for a single image on a classic CPU (that is faster than a low cost CPU) and the trained model is more than 500 MB. Also the SqueezeNet, which is optimized for embedded systems with GPU, is not able to process images in real-time on classic CPUs and on low cost embedded systems. Therefore, the CNNs are not suitable for the purposes of this research work.
The challenge addressed in this thesis is the search of the best trade off between gender recognition accuracy and processing speed, in order to design an algorithm suitable for real-time elaboration on low cost embedded devices. To this aim, the contribution of the thesis is three-fold.

First, an extensive evaluation of various type of classifiers for recognizing the gender on still images is provided. In particular, two multi-experts have been proposed. The former relies upon a fusion of handcrafted features and takes its decision by using information about intensity, texture and shape of a human face. The latter is based on a combination of classifiers fed with trainable shape features and domain-specific descriptors extracted in correspondence of facial landmarks. The multi-experts have been preferred to the simple fusion of different feature vectors in order to reduce the dimensionality, preventing the necessity to increase the amount of data for avoiding over-specialization, and to exploit the complementarity of particular types of features.
All the experiments, aimed at evaluating the accuracy of the proposed methods and at finding the best combination rule, are carried out on standard benchmarks (GENDER-FERET [32] and LFW [21]) with standard protocols. Such preliminary analysis is performed with the goal of optimizing the performance on still images, defining the strengths and the weaknesses of each classifier and the computational load needed to recognize the gender.
Second, differently from the state of the art methodologies, classic surveillance videos instead than high resolution still images have been used for recognizing the gender of the persons. A new dataset has been acquired in real environments and a part of it has been made publicly available for research purposes. Such dataset has been used for learning a classifier able to determine the gender of a person even in presence of challenges such as motion blur, pose variations and so on.
Third, an embedded vision system for real-time gender recognition based on a multi-sensors architecture has been proposed. The choice of a multi-sensors architecture is the best trade-off between accuracy and processing time needed for an embedded vision system, as demonstrated in the experimental evaluation. Such architecture consists of two smart cameras and a low cost embedded system. On board of the first camera, mounted overhead, an efficient and effective people counting algorithm provides the information about the passage of one or more persons and the lo-

cations where the faces can be detected. The information about the position is crucial for reducing the size of the region where the face detection is performed and, consequently, for speeding up the gender recognition algorithm. The second camera, installed in front of the people so that they move towards it, receives the notifications by the previous one and provides the images associated to the passage, together with the information collected by the people counting, to the low cost embedded system devoted to face analysis. The gender recognition algorithm tracks the person in the scene and associates a single classification for each individual, making use of a multi-frame face tracking algorithm. Both the people counting and the gender recognition algorithms are optimized to process images in real-time with very limited resources, thanks to the usage of SIMD instructions which exploit the parallelism of the processor.

1.7 Organization

This thesis consists of the following six chapters:

- **Chapter 1** introduced the topic of gender recognition from still images and videos, by analyzing the challenges and the lack of standard benchmarks, and briefly highlighting the novelties and the contribution of this research.

- **Chapter 2** provides a survey of the state of the art methods, by investigating the approaches recently proposed for recognizing the gender on still images (Section 2.1) and videos (Section 2.2). The chapter ends with Section (2.3), which reports a review of the research in the field of embedded vision.

- **Chapter 3** describes the approaches proposed for gender recognition on still images. In particular, Section 3.1 explains the procedure used for detecting and normalizing the

face images. Section 3.2 describes a method based on the combination of handcrafted features, which takes into account the pixels intensity, the texture and the shape of the human face. Section 3.3 describes the multi-expert which combines trainable COSFIRE filters, namely shape features configured by using prototypes obtained from different parts of the face, and domain-specific SURF-based descriptors, and allows to maximize the gender recognition accuracy on different datasets. Finally, Section 3.4 gives details about the combination rules proposed to optimize the performance of the multi-experts.

- **Chapter 4** is devoted to analyze the multi-sensors architecture proposed for recognizing the gender on real-time video streams. Section 4.2 describes the people counting algorithm used to detect the presence of persons in the scene. while Section 4.3 details the fast and effective gender recognition solution and the tracking algorithm used for associating a single classification to each person. The chapter ends with Section 4.4, where the code optimizations which allow to run the algorithm on embedded systems with limited resources in real-time are discussed.

- **Chapter 5** analyzes the results obtained by the proposed algorithms, both in terms of accuracy and processing speed. Section 5.1 describes the datasets used for the experiments. Sections 5.2 and 5.3 give details about the experimental evaluation of the accuracy achieved by the proposed methods on still images and video sequences, respectively. The Chapter ends with an extensive analysis (Section 5.4) of the processing time required by the proposed approaches, which allows to justify all the choices carried out for the usability of the new architecture in real applications.

- **Chapter 6** draws the conclusions, giving a brief summary of the thesis in Section 6.1 and describing future directions of the research in this field in Section 6.2.

Chapter 2

State of the Art

2.1 Gender recognition on still images

Although it is not possible to define a taxonomy to partition the methods for gender recognition from still images, two different classes can be roughly recognized, depending on the type of features used. Most of the approaches rely on *handcrafted* features, which require expert knowledge for manually designing domain-specific features. Other approaches are indeed *trainable*, in that distinctive features can be automatically learned from training data. The advantage of using handcrafted features is the possibility to exploit the domain knowledge to identify the elements that distinguish the faces of men from those of women, such as intensity, texture, shape and geometry. Indeed, trainable features may capture aspects of the face that a human could not notice. Moreover the procedure for the extraction of such features does not rely upon domain knowledge.
The approaches belonging to the first category use various types of features based on color [33] [34] [35], texture [36] [37] [38] and shape [39] [40] [41] information. Almost all of them share a similar architecture that consists of three steps: i) the detection and the cropping of the face using the well known Viola-Jones algorithm [42]; ii) the pre-processing of the image, in order to normalize the face in terms of dimension, pose and illumination; iii) the extrac-

tion of the features used to recognize the gender.
For example, in [33] Moghaddam et al. propose to use raw information (the pixel intensity values of face images) to form vectors and use them to train a SVM classifier with an RBF or a polynomial kernel. Their main drawback is that they are not invariant to translation. If the same face is shifted by just a few pixels, the resulting feature vector may be completely different. Another problem is the dimensionality of the obtained feature vector, which increases with the resolution of the face image. In order to address this problem, Yang et al. [34] compared the performance of various dimensionality reduction techniques applied on the pixel intensity values, such as principal component analysis (PCA), 2D principal component analysis (2DPCA), independent component analysis (ICA) and linear discriminant analysis (LDA). In [35] Baluja et al. use the relationship between the intensities of the pixels in face images. They consider ten types of a pixel comparison operator, which provide a binary decision, as weak classifiers to learn a model using the Adaboost method [43].
Lian et al. [36] extract and concatenate local binary pattern (LBP) histograms [44], from different regions of the face, in a single feature vector, and trained a SVM classifier for gender recognition. The rationale of this approach is that a texture descriptor could be able to capture the differences between the smoother skin of a woman and the rougher skin of a man, especially in presence of beard. Eidinger et al. [18] and Azarmehr et al. [19] use a pair of different LBP variants, FBLBP and MSLBP respectively, for the automatic recognition of gender and age. Dago-Casas et al. [45] extract Gabor wavelets from a sparse uniform grid of face points and compute a face descriptor combining them with LBP histograms. Since not all the regions of the face are significant in terms of texture, in [37] and [38] other researchers propose to use Adaboost to carry out feature selection and to use only the LBP histograms from the most discriminant parts of the faces.
In [39] Singh et al. propose the use of histograms of gradients (HOG) [46] to represent the shape of a face and use it as a descriptor for gender recognition. In [40] Guo et al. demonstrate that

the performance of a gender classifier based on HOG features is affected by age. This idea is further investigated in [41], where the authors find dependencies among facial demographic attributes, especially between gender, age and pose facial attributes. Other researchers also try to combine several typologies of color, shape and texture features, in order to improve the performance of their gender classifiers [47] [48]. The rationale behind those approaches is that color, texture and shape features can be complementary, in the sense that they capture different aspects of human faces and can improve gender recognition when used together.
Other domain specific approaches rely on the extraction of handcrafted features from specific points, known as fiducial points [49]. Brunelli et al. [50] propose a face descriptor that computes 18 fiducial distances between points representing the locations of the eyes, nose, chin, mouth corners and others. El-Din et al. [51] extract SIFT [52] descriptors from these so-called facial landmarks and used them to form a long feature vector.
As for the second category, the deep learning-based methods [53] [54] [55] [56], which gained popularity in recent years, are the most common trainable approaches. Levi et al. [53] perform automatic age and gender classification using deep-convolutional neural networks (CNN) [57]. Van de Wolfshaar et al. [54] train a dropout-SVM using the deep features selected by a CNN. Ranjan et al. [55] propose a multi-task learning framework, called HyperFace, for simultaneous face detection, landmark localization, pose estimation and gender recognition using CNNs. Jia et al. [56] design a method to generate a classifier using a mixture of face datasets composed of four million images and about $60,000$ features.
Although the deep networks have the capability to achieve very high gender recognition accuracy, they require the use of a GPU and are not able to process images in real-time on classic CPUs, much less on low cost embedded systems. Moreover, the CNN trained model requires a lot of memory (more than 500 MB) that is often not available on the most popular embedded devices for computer vision, namely the smart cameras.

2.2 Gender recognition on videos

The gender recognition algorithms proposed in the literature deal with scenarios where the training and the test samples are acquired in similar environmental conditions. In most of the cases the environment is controlled and the quality of the images is very high. Obviously such face images are very different from the ones acquired with classic video-surveillance cameras. Experiments performed in real environments [11] demonstrate that these approaches fail to achieve an acceptable accuracy when dealing with images acquired in different environmental conditions, because training and test samples differ in resolution, contrast and sharpness. The drawbacks of these techniques are: i) the lack of experiments with cameras installed in real scenarios and ii) the inability to cope with the changes in the distributions of training and testing conditions.
As for the first point, the most challenging datasets where the recent approaches are achieving remarkable results are the LFW [21], the Adience [18] and the YoutubeFaces [23]. However, both the datasets are acquired with professional cameras or, anyway, with devices which assure performance better than the classic surveillance cameras. Datasets composed of images acquired in real scenarios have not yet been acquired for gender recognition purposes. Nevertheless, the benchmarks SCFace [30], ChokePoint [29] and FR_SURV_VID [31] have been used for addressing the problem of face recognition in surveillance scenarios. The latter problem is strictly related to the second point of the discussion, since in most of the surveillance systems just a few face images of the person (generally taken from ID card or driving license) are available in a database and it is very hard to recognize or verify the identity of a person by comparing the reference faces with the video images captured with the surveillance cameras.
The challenge of the systems which aim to address this problem is to find features that are invariant to the degradation of the image quality and to the well known face variations in real environments. Although the problem is very interesting and challenging also for

gender recognition purposes, no researches have been conducted in this field, but only for face recognition in surveillance scenarios. The research in the field of face recognition in real environments aims substantially at finding a suitable domain adaptation task, able to minimize the discrepancy between the probability distributions of the training and testing domains. Domain adaptation is a fundamental problem in machine learning [58, 59, 60, 61, 62, 63, 64, 65] and has gained a lot of traction in natural language processing, statistics, machine learning, and, most recently, in the computer vision field [66, 67]. The most common usage of domain adaptation in the field of face recognition with surveillance cameras is to solve the single sample per person (SSPP) problem [68]. Indeed, in some specific scenarios (e.g. law enforcement, driver license, passport and identification card) only one image per person can be acquired for the training of face recognition systems, so making even harder the estimation of the intra-class variation with only a single training sample. To deal with such challenging problem, several pattern recognition based or data driven approaches have been proposed.

The methods based on pattern recognition address the problem by predicting all the possible face variations, in order to adapt the model of the face to the real environment [69, 70, 71, 72, 73, 74]. The authors of the papers [69, 70, 71] concentrate their attention on the training step, proposing, respectively, the learning of generic discriminant vectors, an adaptive linear regression and an incremental learning. The authors of [72] construct a local gallery dictionary by extracting the neighbouring patches from the gallery dataset and an intra-class variation dictionary by using an external generic dataset for predicting the possible facial variations (e.g., illuminations, pose, expressions and disguises). The algorithm takes the advantages of patch based local representation and generic variation representation to deal with SSPP. A similar approach has been used in [73], where a sparse variation dictionary learning has been proposed. In [74] a reference face graph has been designed for addressing the SSPP problem.

The data driven approaches address the SSPP problem by pro-

ducing virtual training samples from the real samples. The idea is that to learn the intra-class variation more samples are needed and that additional face images of the same person may be obtained as degraded versions of the original training sample with geometric transformations [75, 76, 77] or specific image processing [78, 79, 80, 81, 82]. In [75] the original face image is rotated using bilinear interpolation to form more training faces for every person. The authors of [76] propose sampled FLDA to partition the single face into several sub-image by sampling interval in height and width respectively. Similarly in [77] the training sample has been divided transversely into several parts and a sparse representation has been generated by combining original and virtual samples. The authors in [78] use multi-directional orthogonal gradient phase faces to handle illumination invariant single sample face recognition. In order to constitute a variational feature representation from single sample per person, in [79] a linear regression model to fit the variational information of a non-ideal probe sample with respect to an ideal gallery sample has been proposed. 3D face reconstruction [80] has become an effective tool to deal with SSPP in recent years. In [81] a personalized 3D face model is firstly constructed from a single frontal 2D face image with neutral expression and normal illumination and then realistic virtual faces with different pose, illumination and expression are synthesized. Finally, the authors in [82] use lower-upper decomposition algorithm to decompose single sample into two basis images set in order to reconstruct two approximation images from the two basis.

Most of the above mentioned approaches aim at optimizing just a single parameter, namely the recognition accuracy. However, only a modest attention has been devoted to the computational optimization of this kind of algorithms and to the analysis of the required hardware resources which allow to use these systems in real time and to apply them in real applications. Furthermore, as confirmed in [83], only a few fast implementations of gender recognition algorithms have been proposed in the recent years. For instance in [19] an Enhanced Discriminant Analysis (EDA) is

combined with a SVM classifier based on RBF kernel exploiting a demography-based discriminative model. In [83] a Neural Network based on a Gabor filter is implemented over a FPGA, while in [84] the distances between facial features are evaluated over a FPGA.
Other fast face analysis solutions have been proposed for face recognition in real-time, by optimizing the software [85, 86] or directly the hardware [87]. Since the most time requiring operation is the face detection, various approaches to speed up this processing step have been also proposed [88, 89].
However, in all the above mentioned methods, no experimentations have been made over sequences of images acquired in real scenarios.

2.3 Embedded vision

The Embedded Vision Alliance [90] defines the new discipline of embedded vision as the practical use of computer vision in machines that understand their environment through visual means. The terms embedded refers to the use of digital processing and intelligent algorithms to interpret meaning from images or video on low-cost and energy-efficient processors.
The research in the field of embedded vision aims substantially at achieving two goals. On one hand, the finding of special purpose hardware solutions which assure performance, in terms of processing time, comparable with the general purpose architectures based on graphic processing unit (GPU). On the other hand, the effort of the researchers is concentrated on the design of algorithms which are able to exploit the features of such architectures and to perform the specific computer vision task in real-time, by finding the best trade-off between accuracy and processing time.
As for the hardware architectures, the best candidates are the Application Specific Integrated Circuit (ASIC), Field Programmable Gate Array (FPGA) and System on a Chip (SoC). The suitability of FPGA for computer vision applications is discussed in [91]. In

[92] it is shown that a special purpose FPGA network can achieve a performance comparable with a GPU, but with a significantly lower energy consumption. The performance of FPGA, CPU and GPU are also compared on three different image processing applications in [93]. Some FPGA implementations of the Speeded Up Robust Features (SURF) algorithm have been proposed [94, 95]. An improved version of the SURF algorithm, which combine the use of FPGA and SoC, is described in [96]. Other researchers proposed their own architectures based on domain specific accelerators [97], ultra-low power accelerators [98], parallel pipelined heterogeneous SoC [99] and 3D cameras [100].
The SoC architectures are very interesting because generally consists of an ARM-based CPU which can be extended with a FPGA, a GPU or a Digital Signal Processor (DSP) in order to achieve performance comparable with the ones obtained by the modern processors, without increasing the cost and the energy consumption. The smart cameras, which are probably the most used embedded vision architectures, belong to this class. Most of the low level algorithms have been optimized for such devices. In [101] it is described a background subtraction method that does not perform floating point operations and largely uses SIMD (Single Input Multiple Data) instructions, thus exploiting the parallelism of the vector processors and simultaneously processing more data. Since integral images are largely used as preliminary operations for various feature extraction techniques, efficient algorithms for their computation on embedded vision systems are presented in [102]. An optimized implementation of the HOG algorithm, suitable for embedded systems, is reported in [103].
The trend of embedded vision has led to a proliferation of industrial, automotive, security and retail applications based on these architectures. For example, a bio-inspired embedded vision system for autonomous micro-robots is proposed in [104] to improve the factory automation. A reconfigurable embedded vision system for advanced driver assistance, which ensure that the driver remains alert and awake while driving the vehicle, is presented in [105]. In [106] an algorithm for detecting abandoned baggages or removed

objects by using surveillance smart cameras in real-time is proposed. In [107] an embedded vision method for fast notification of vehicles parked in forbidden areas is described. A very fast and effective method for counting people on real video sequences by using an overhead smart camera is proposed in [108].
Although significant interest has been registered in the last years, the research in this field is yet at the beginning. The challenge of the researchers is the definition of algorithms as fast as accurate by using the very limited resources available on smart cameras or other embedded vision systems.

Chapter 3

Gender recognition on still images

Going back to the considerations about the pros and cons of using domain-specific or trainable features for gender recognition, in this research both the approaches have been investigated.
Up to now the efforts of the research community have been mainly devoted to the definition of a representation of face able to discriminate men from women in all the conditions. Considering the amount of possible variations in terms of age, race, pose, illumination, occlusions and so on, a description which takes into account all these situations inevitably leads to high dimensional feature vectors.
How to manage high dimensional feature vectors is a well-known problem in the communities of machine learning and pattern recognition: in fact, as shown in [109], employing a high dimensional feature vector would imply significant increase in the amount of data required to train any classifier in order to avoid over-specialization and to achieve good results. Furthermore, independently of the particular features extracted, in most of the above mentioned methods the high variability of faces, as well as the large amount of noise in data acquired in real environments, prevent the systems from the achievement of a high recognition rate. More generally, it has been shown [110] that increasing the performance of a system

based on the traditional combination feature vector - classifier is often a very expensive operation. In fact, it may require to design a new set of features to represent the faces, to train again the classifier, or to select a different classifier if the performance is not sufficiently satisfactory. Moreover, this effort could be payed back by only a slight improvement in the overall accuracy, so this approach may prove not very convenient.
In order to overcome the above mentioned limitations, one of the solutions coming from the literature [110] is to split the feature vector and consequently to adopt a set of classifiers, each tailored on a feature set and then trained to be an expert in a part of the feature space. The main idea of this kind of paradigm, usually referred to as Multi Expert System (MES), is to make the decision by combining the opinions of the different individual classifiers (hereinafter experts), so as to consistently outperform the single best classifier [111]. Such research explains on the basis of a theoretical framework why a MES can be expected to outperform a single, monolithic classifier. In fact, most classifiers, given an unlimited amount of training data, converge to an optimal classification decision (in a probabilistic sense); but on a finite training set, their output is affected by an error (additional with respect to the inherent error due to ambiguities in the input data), which is either due to over-specialization or to the choice of reducing the classifier complexity in order to avoid the loss of generalization ability. The author of [111] shows that, under some assumptions satisfied very often in practical cases, a suitably chosen benevolent combining function can make the overall output of the MES less sensitive to the errors of the individual classifiers. MESs have been successfully applied in several application domains, ranging from biomedical images analysis [112] [113] and face detection [114] to movie segmentation [115] and handwriting recognition [110].
In the proposed gender recognition methods, which work on still images, MESs are employed for improving the performance of the single experts. It is evident that the successful implementation of a MES requires both the adoption of complementary sets of features feeding the different experts and the choice of a reliable

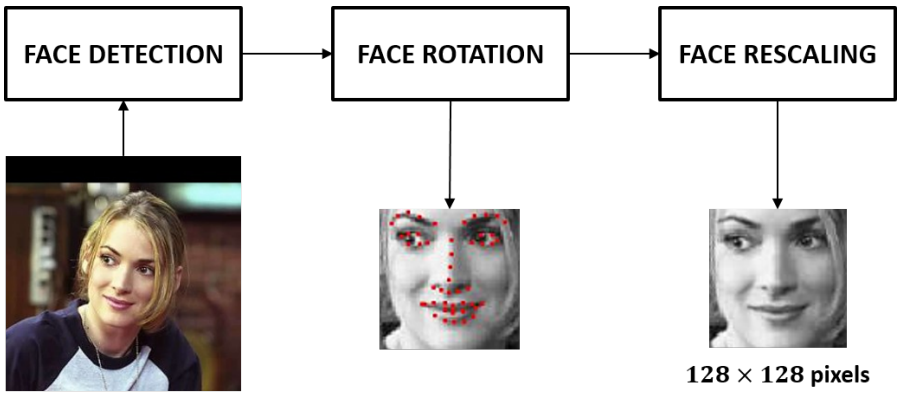


Figure 3.1 Overview of the face detection and normalization algorithm.

combination rule.
Regarding the first aspect, two multi-experts have been proposed. The former, described in Section 3.2, considers three different experts able to analyze the problem of gender recognition from face images from different points of view, based on color intensity, on texture and on shape, respectively. The latter, reported in Section 3.3, combines a classifier tailored with trainable COSFIRE filters, which allow to design a face descriptor able to automatically capture particular facial features, with an expert based on domain-specific SURF features extracted from specific facial landmarks.
As for the second aspect, an experimental evaluation has been performed to choose the best combination rule for the problem at hand. In particular, a weighted voting rule has been used for combining the decisions of the classifiers based on handcrafted features, while a data driven stacked classification scheme has been used for the combination of trainable and domain-specific features.

3.1 Face detection and normalization

Figure 3.1 illustrates the architecture of the proposed face detection and normalization algorithm.

Preliminarily, the Viola-Jones algorithm [42] is applied on the

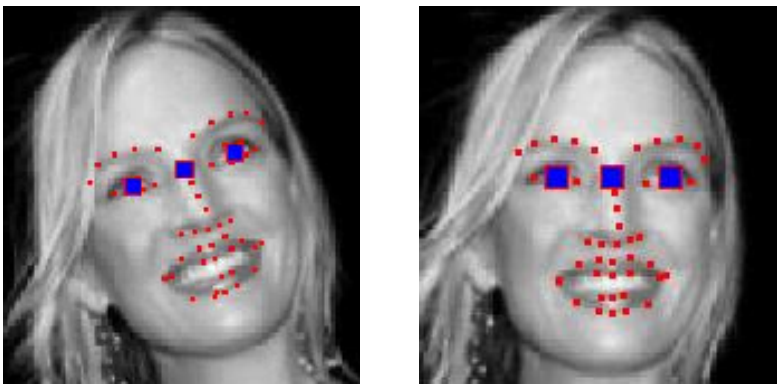


Figure 3.2 Representation of the proposed face alignment algorithm. The 51 red dots indicate the positions of the facial landmarks. The three blue markers, from left to right, indicate the left eye center, the center of the line that connects the two eyes, and the right eye center.

input image to detect the faces occurring in it. Then a face alignment algorithm is used to normalize the pose. For a given face image, the method proposed in [116] is used to detect a set of 51 facial landmarks. The average location of each of the two sets of eye-related landmarks is computed. This allows to determine the orientation of the line which connects these two points, namely the orientation of the face, and use that angle to horizontally align the face image. In order to horizontally align the face, the image is rotated around the center of the line that connects the two eyes. Figure 3.2 depicts an example of a face image before and after the alignment. In practice, in order to avoid having a black background in the rotated image, first the face image is cropped by using a bounding box that is twice as large as the one determined by the Viola-Jones algorithm. Then the image is rotated and the Viola-Jones bounding box is used to crop the face in the rotated image. Finally, each horizontally aligned face image is rescaled to a fixed size of 128×128 pixels. Such resolution allows to maximize the accuracy of the experts described in the following sections.

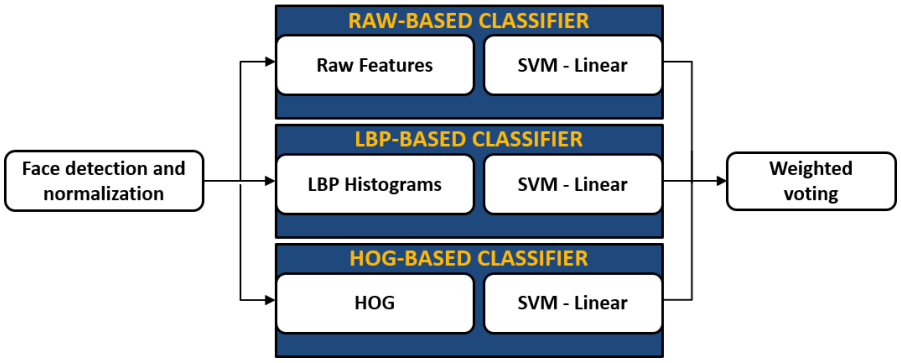


Figure 3.3 Architecture of the method based on a fusion of handcrafted features.

3.2 Gender recognition on still images using a fusion of handcrafted features

Figure 3.3 shows the architecture of the method which rely on the fusion of raw, texture and shape features. The algorithm for face detection and normalization has been described in Section 3.1. The experts based on handcrafted features are detailed in the following subsections, while the weighted voting rule is described in Section 3.4.

3.2.1 Raw-based classifier

As for the raw-based descriptor, the image is rescaled to 64×64 pixels and transformed into a $(64 \times 64 =)$ 4096-element feature vector, dividing each element by 255 so that all dimensions have the same range of [0,1]. Such descriptor has been used to feed an SVM classifier with a linear kernel.

3.2.2 LBP-based classifier

The texture feature vector is obtained by applying the LBP descriptor [44] to the entire image and comparing the intensity value

of each pixel with a 3×3 neighbourhood. A spatial tiling of 3×3 is used to generate a 256-element L2-normalized histogram for each tile. Finally, the nine histograms are merged so as to form a ($256 \times 9 =$) 2304-element vector for each image. The LBP histogram-based descriptors are used as inputs of an SVM classifier with a linear kernel.

3.2.3 HOG-based classifier

As to the shape descriptor, firstly the gradient and angle of every pixel are computed by considering the responses of first-order partial derivatives of a 2D Gaussian function with a $\sigma = 1$. Then blocks of 32×32 pixels that overlap by 50% are sampled and for each block a spatial tiling of 2×2 is used. For each tile the L2-normalized weighted histogram of 9 bins (in intervals of 20 degrees) is computed, the normalized values are clipped at 0.2 and all the values are normalized again. Considering that face images of size 128×128 pixels are used, the HOG descriptor results in a (7 blocks $\times$ 7 blocks $\times$ 4 tiles $\times$ 9 bins =) 1764-element vector. Also in this case, an SVM classifier with a linear kernel has been trained with the above mentioned shape features.

3.3 Gender recognition on still images using a fusion of domain-specific and trainable COSFIRE filters

The idea of this approach is to combine the domain-free and trainable COSFIRE filters, configured using aligned face images, with the handcrafted SURF features [117] extracted from 51 facial landmarks related to eyes, nose and mouth. Hereinafter these methods will be named *COSFIRE-based* and *SURF-based*, respectively.
The expectation is that such features are complementary since they capture, in principle, different aspects of the human face.

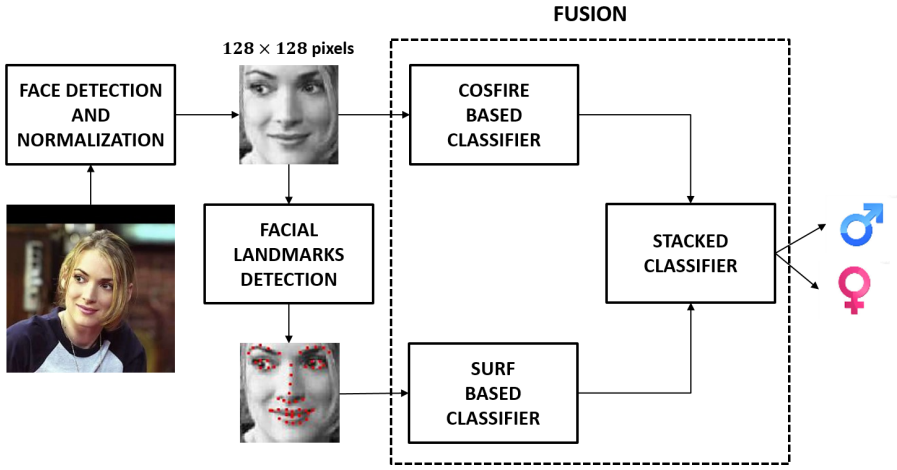


Figure 3.4 Architecture of the method based on the fusion of domain-specific and trainable features.

The COSFIRE-based classifier should be able to find the differences in the shape of male and female faces, while the SURF-based classifier should rely on the descriptors extracted from the facial landmarks to discriminate the local differences between the faces of men and women. For these reasons the proposed method should increase the robustness to the aforementioned face variations. Moreover, the fusion of trainable and handcrafted features should ensure a better tradeoff between generalization and specificity. The proposed approach is the first one that combines domain-independent features with domain-specific ones.

3.3.1 COSFIRE-based classifier

Trainable COSFIRE filters have been demonstrated to be effective in various computer vision applications, including object localization and recognition [118, 119, 120], vessel-like segmentation [121, 122], and contour detection [123, 124].
A COSFIRE filter is trainable, in that its selectivity is determined in a one-step configuration process that automatically analyzes a given prototype pattern of interest. The resulting non-linear COSFIRE filter can then be applied to images in order to localize

patterns that, to a certain extent, are similar to the prototype. Below it is briefly described the required processing to configure and apply COSFIRE filters, and subsequently use their responses to form feature vectors.

3.3.1.1 Configuration of a COSFIRE filter

The idea of a COSFIRE filter is to combine the responses of some low-level detectors that are selective for simple features in order to determine the selectivity for a more complex feature or a shape. In [118], for instance, it was shown that by combining the responses of orientation-selective Gabor filters at certain positions, one could configure a COSFIRE filter that is selective for a complex shape, such as a traffic sign. By simply changing the input low-level detectors from Gabor filters to difference-of-Gaussians, one could achieve very effective contour [123, 124] and vessel [121, 122] detectors. These two types of COSFIRE filters are essentially shape detectors and do not take colour into account. Recently, a new type of COSFIRE filters were proposed which take input from color blob detectors and have been found to be more effective than Gabor-based COSFIRE filters in object recognition datasets where colour plays an important role [120].
In this method the original Gabor-based type of COSFIRE filters [118] are used, as colour is not considered to be a distinctive feature for gender recognition. In an automatic configuration process a bank of Gabor filters with eight orientations and five scales is firstly applied and their responses are superimposed. Then a number of concentric circles around a point of interest are considered and the positions along these circles with local maxima Gabor responses are determined. For each such a point a tuple with four parameters $(\lambda, \theta, \rho, \phi)$ is formed, where λ nd θ denote the scale and orientation, respectively, of the Gabor filter that achieves the maximum response at that position, while ρ and ϕ, respectively, denote the distance and polar angle with regards to the point of interest. Finally, $S_f = \{(\lambda_i, \theta_i, \rho_i, \phi_i) \mid i \in 1 \dots n\}$ denotes a set that contains the 4-tuples that represent all n points at which lo-

cal maximum Gabor responses are achieved.
The center point used in a given prototype is the position at which the resulting COSFIRE filter will obtain the maximum response. It can either be specified manually or selected automatically. For the application at hand, such locations are chosen randomly in the training face images and their surroundings are used as local prototype patterns to configure COSFIRE filters. In this way, a COSFIRE filter is selective for a small part of a face.
Figure 3.5 shows the configuration procedure of two COSFIRE filters by using parts of the eyebrows as prototype patterns selected from a male and a female face images.

3.3.1.2 Response of a Gabor-based COSFIRE filter

The response of a Gabor-based COSFIRE filter is computed in four simple steps, namely filter-blur-shift-multiply. In the first step the unique pairs of the parameters (λ, θ) from the set S_f is determined and Gabor linear filtering in the Fourier domain with those parameter values is applied. Secondly, in order to allow for some tolerance with respect to the preferred positions, for the $i-$th tuple the corresponding Gabor response map is blurred with a Gaussian function whose standard deviation σ_i is a linear function of the distance ρ_i. In practice, the linear function $\sigma_i = \sigma_0 + \alpha\rho_i$, with σ_0 and α set to the default values ($\sigma_0 = 0.67$, $\alpha = 0.1$) proposed in [118] is used. Thirdly, each blurred Gabor response is shifted by the polar vector $(\rho_i, -\phi_i)$, so that all afferent Gabor responses meet at the support center of the concerned COSFIRE filter. Finally, the geometric mean function, essentially multiplication, is used to combine all blurred and shifted Gabor responses and come to a scalar value in each position of a given image. Figure 3.5(d) and Figure 3.5(h) show the response maps of the two COSFIRE filters applied to the two images from which some local patterns are used for their configuration. They respond in locations where the local patterns are very similar to the eyebrow prototype parts.

In [118], it was also demonstrated how tolerance to rotation, scale and reflection could be achieved by the manipulation of pa-

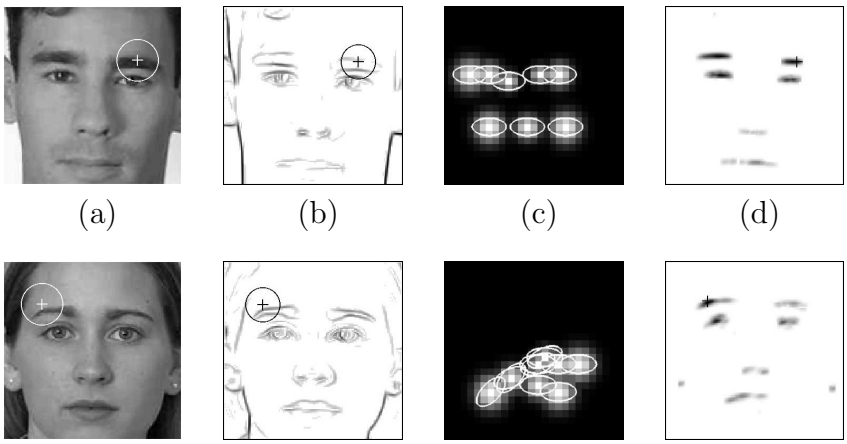


Figure 3.5 Configuration of two COSFIRE filters using (a-d) a training male face image and (e-h) a training female face image, both of size 128×128 pixels. (a,e) The encircled regions indicate the prototype patterns of interest which are used to configure the two COSFIRE filters. (b,f) The superposition of inverted response maps of a bank of Gabor filters with 16 orientations ($\theta = \{0, \pi/8, \ldots 15\pi/8\}$) and a single scale ($\lambda = 4$). (c,g) The structures of the COSFIRE filters that are configured to be selective for the prototype patterns indicated in (a) and (e). (d,h) The inverted response maps of the concerned COSFIRE filters to the input face images in (a) and (e). The darker the pixel the higher the response.

rameter values. These invariances are, however, not necessary for this application.

3.3.1.3 Forming a feature descriptor and learning a classification model

By using k local patterns randomly selected from the training face images, k COSFIRE filters that are selective for different parts of male and female faces are configured. For a given face image the collection of k COSFIRE filters is then applied and a spatial pyramid of three levels from is used to take the COSFIRE filter responses. In level zero, where there is only one tile, the global maximum responses of all COSFIRE filters across the entire image are taken. In level one and level two, each COSFIRE response map

is divided, respectively, into $(2 \times 2 =)$ 4 and $(4 \times 4 =)$ 16 tiles and the maximum response in each tile is taken. For k COSFIRE filters and a spatial pyramid of $(1 + 4 + 16 =)$ 21 tiles the face image is described with a $21k-$element feature vector. The set of k COSFIRE filter maximum responses per tile is then normalized to unit length. Figure 3.6 depicts the spatial pyramids of the two above configured COSFIRE filters obtained from a test female face image and a bar graph with the values of the resulting $(21 \times 2 =)$ 42-elements descriptor.

The $21k-$element feature vectors of all training images are used to train a SVM classification model with the following chi-squared kernel:

$$K(x_i, y_i) = \frac{(x_i - y_j)^2}{\frac{1}{2}(x_i + y_j) + \epsilon} \tag{3.1}$$

where x_i and y_j are the feature vectors of the $i-$th and the $j-$th training images, while the parameter ϵ represents a very small value and it is used to avoid numerical errors. This COSFIRE-based descriptor is inspired by the concept of population coding from neuroscience [125] as well as from the spatial pyramid matching approach [126].

3.3.2 SURF-based classifier

All the pre-processing steps needed to detect and align the face image are the same of that described in Section 3.1. In the SURF-based classifier, the 51 facial landmarks belonging to the eyes, the nose and the mouth are detected on the aligned face and the SURF descriptors at the keypoints indicating the facial landmarks are extracted. This approach results in a $(51 \times 128 =)$ 6528-element feature vector for each face image. Finally, these vectors are used to train a SVM classifier with a linear kernel.

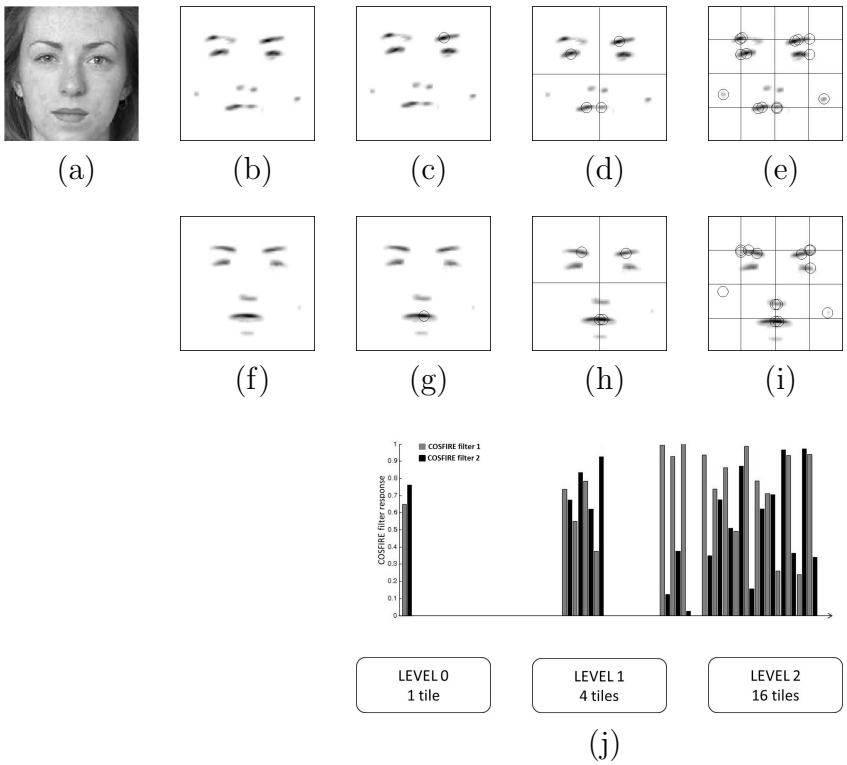


Figure 3.6 Application of the two COSFIRE filters to a test face image. (b,f) Consideration of three-level spatial pyramids to the response maps of the COSFIRE filters. (c,g) In level zero only one tile, which has the same size of the given image, is considered. (d,h) In level one four tiles in a 2×2 spatial arrangement are taken into account. (e,i) In level two16 tiles in a 4×4 grid are considered. For each of the 21 tiles the circle indicates the location of the maximum response. (j) The resulting face descriptor, which consists of $(21 \times 2 =)$ 42 values. The responses are normalized to unit length for each tile.

3.4 Combination rules

As mentioned before, one of the main choices determining the performance of a MES is the combination rule. Although several strategies have been proposed in the last years [127], it has been proved that one of the most robust to errors of the combined classifiers (both when combining classifiers based on the same features

and when using different feature subsets in each classifier) is the majority voting rule [111]. The idea is that each of the experts should express its vote in terms of a pair of probabilities that a given image has a male or a female face. In order to come up with a decision the three male probabilities and the three female probabilities should be summed and if the total male probability is greater than the total female probability, then the given face image is labelled to be a male otherwise a female.
A variation of the majority voting is the weighted voting rule [111], where the votes of the experts are weighted proportionally to the recognition rate achieved for each class on the training set. However, this technique assigns an absolute weight to each expert, without considering particular situations where a classifier that is generally less effective may be more reliable.
In order to learn from the data the best way to combine the classifiers, the decisions made by the experts proposed in this chapter are combined using a stacked classification scheme. This technique learns the combiner algorithm by training a new classifier with all the predictions of the single experts. It can be considered a very smart variation of the weighted voting rule [111]. The difference is that, in case of the stacked classification scheme, the regions of decision, instead of the weights, are learned during the training.
After an experimental evaluation (see Section 5.2.5 for more details), the best combination rule for the fusion of handcrafted features is the majority voting. The same experimental analysis points out that the more effective combination rule for the COSFIRE- and the SURF-based classifiers is the stacked classification scheme. So, in this case, the output scores of the single SVM classifiers achieved from the training images are used as feature vectors to learn another SVM with a linear kernel. The final layer determines the classification of a given test image.

Chapter 4

Gender recognition on real-time video streams

In the previous Chapter various methods for gender recognition on still images have been proposed. The purpose of that approaches is to maximize the gender recognition accuracy, by finding the features and the classifiers able to describe the human face in the best way for recognizing the gender. Section 5.2 will demonstrate the effectiveness of the proposed methods on the considered datasets. However, as reported in the Introduction, additional challenges have to be taken into account dealing with video sequences acquired in real scenarios. Section 5.3 will show that the approach described below is able to achieve a remarkable gender recognition accuracy, thanks to a data-driven learning carried out with a new dataset of images acquired in real scenarios which contains a wide range of face variations. The most important contribution, nevertheless, is the solution provided for allowing the real-time gender recognition on low cost embedded systems.
The proposed embedded vision architecture derives from an extensive analysis of the computational load, provided in Section 5.4, which gives essential insights for finding the best trade-off between accuracy and processing speed. Indeed, three fundamental choices have been performed after the computational requirements analysis.

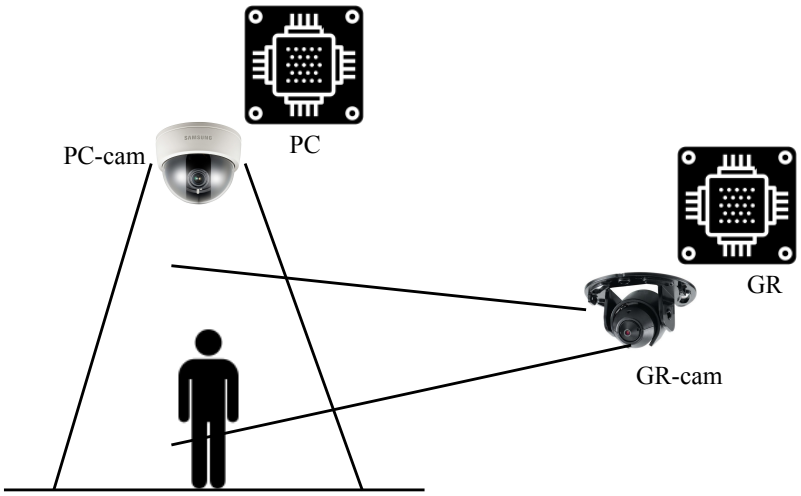
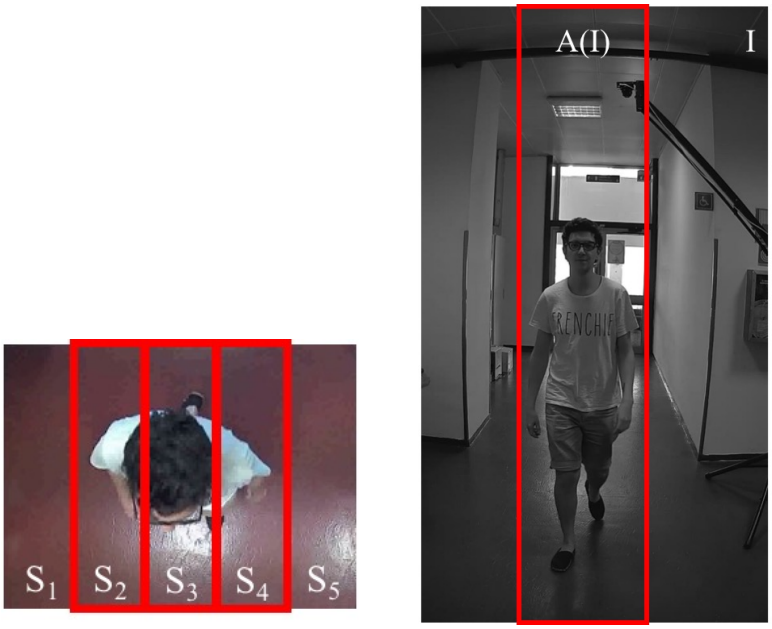


Figure 4.1 Overview of the proposed architecture. *PC* and *GR* are the modules in charge of counting people and recognizing the gender of the persons, respectively.

First, the face alignment algorithm is a very costly operation, mainly due to computational effort required by the facial landmarks detection. This step has been consequently removed from the method described below.
Second, the HOG-based classifier has been selected as the best trade-off between accuracy and processing speed. Indeed, the SURF-based classifier can not be used for the same considerations pointed out for the face alignment, while the COSFIRE-based method requires more than one second for the computation of the descriptor. The RAW-based classifier is not effective with faces that are not aligned, since a small variation in the pose may completely change the face descriptor and, consequently, the result of the classification. Section 5.2 will highlight that the HOG-based classifier is able to achieve the best accuracy if considered as single expert and that the combination with the LBP-based expert does not improve the accuracy.
Third, also performing the best optimization of the gender recognition solution, it is not able alone to carry our real-time analysis



(a) PC-cam (b) GR-cam

Figure 4.2 View of the two cameras, respectively PC-cam (a) and GR-cam (b), with the counting sensor (a) and the area to analyze for the gender recognition (b) overlapped on the images.

on a single low cost embedded system. For this reason, a multi-sensor embedded vision architecture composed of a smart camera devoted to people counting and a couple camera-low cost device dedicated to gender recognition. The details about the proposed architecture are provided in the following.

4.1 Overview

The proposed approach is based on a distributed and multi sensors architecture; the main idea lies in the fact that one of the most burdensome steps of gender recognition algorithms is the face detection. However, it is not strictly necessary to perform this step for each frame and in the whole image, but it could be

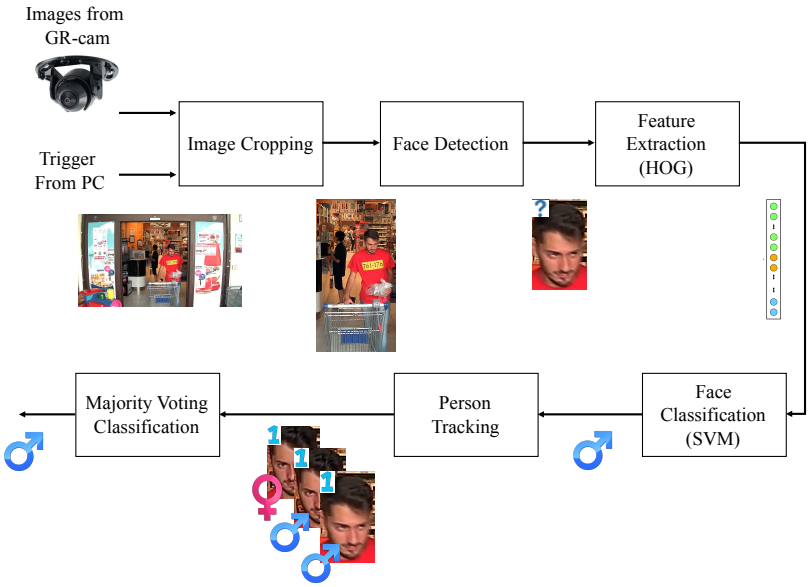


Figure 4.3 Overview of the proposed gender recognition algorithm.

ideally performed only in those situations where there are surely one or more persons taken by the camera and in the regions where a face may be detected. Starting from these observations, the idea is to combine a head-view people counting module with a gender recognition one. An overview of the proposed approach is shown in Figure 4.1. The system consists of two different modules: the module in charge of counting people (hereinafter People Counting module, *PC*) which analyzes the video stream acquired by a head-mounted camera (hereinafter *PC-cam*) and the module in charge of recognizing the gender (hereinafter Gender Recognition module, *GR*) of the persons moving towards a frontal-view camera (hereinafter *GR-cam*). An example of images acquired by PC-cam and GR-cam, respectively, are reported in Figure 4.2.

In more details, PC is always active and has the responsibility to detect the passage of a person through a virtual line. As soon as an event related to the passage of a person under the PC-cam is detected by PC, a trigger to GR is sent and the search of faces

can start in a limited set of images (namely N before and N after the trigger). This search is not performed over the whole image, but just over a limited portion of the scene, corresponding to the area where the passage of the person has been detected.

The main advantage deriving from the proposed approach lies in the fact that the computational burden of the gender recognition module is strongly reduced since only a low percentage of frames is processed. Such optimized architecture allows to run both the algorithms (namely, PC and GR), over SoCs, thus strongly reducing the cost of the hardware infrastructure and making possible its usage in real systems. It is important to highlight that the computational improvement is not paid in terms of accuracy, as shown in Chapter 5.

More details concerning the algorithms will be provided in the following: in Section 4.2 the people counting algorithm will be presented, while in Section 4.3 a description of the proposed approach for recognizing the gender will be detailed.

4.2 People Counting

In order to maximize the performance of the face detection, it is important to capture all the images related to a passage, namely all the frames that are good candidates for containing a face. For this reason it has been adopted the people counting algorithm recently proposed in [108], due to its efficiency and its robustness. The main idea of this algorithm is to use the foreground mask, extracted with a traditional background subtraction and updating algorithm, so as to feed a *virtual sensor* which works like an incremental rotary encoder. The sensor, as shown in Figure 4.4, is characterized by a rectangular area with a crossing direction. The width W of the sensor is the side perpendicular to the crossing direction, identified in the figure by the arrow on the right), while the height of the sensor H is the side parallel to the crossing direction. The sensor is partitioned into K different stripes S_i, with $i = \{1, ..., K\}$, each of one having a width equal to $\delta = W/K$. The

i-th stripe S_i is composed by two adjacent cells C_{ij}, with $j = 1, 2$. A cell is active if a sufficient percentage p_c of foreground pixels in that cell $F(C_{ij})$ with respect to the area of the cell $Area(C_{ij})$ is present. The state c_{ij} (being $c_{ij} \in \{0, 1\}$) of the cell C_{ij} can be computed as follows:

$$c_{ij} = \begin{cases} 1, & \text{if } F(C_{ij})/Area(C_{ij}) > p_c \\ 0, & \text{otherwise} \end{cases} \tag{4.1}$$

Anyway, the algorithm does not consider only the information related to the current time instant and thus to the state activation c_{ij}. Indeed, it also stores the previous state c'_{ij} in order to evaluate the activation of the corresponding stripe S_i. In more details, S_i can be considered activated (then it is equal to 1) if the following condition holds:

$$c_{i2} = 1 \wedge c'_{i1} = 1 \wedge c'_{i2} = 0 \tag{4.2}$$

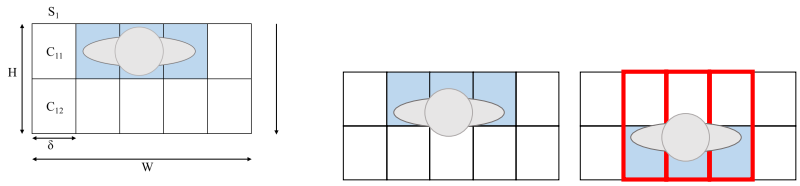
In order to identify the passage of a person, the algorithm verifies the relative position of the stripes. For each set of L adjacent stripes, a person is counted. As evident, the different sets of stripes accounting for a passage can not contain any common stripe. Thus, the set of active stripes AS contributing to a counting is the following:

$$AS = \bigcup_i S_i, ..., S_{i+L} \tag{4.3}$$

with $1 \leq i \leq K$.

In the experiments the parameters have been set as follows:

- L has been experimentally set to 3, as suggested in [108]; indeed, this value guarantees the best possible tradeoff between the accuracy that can be achieved by the proposed approach and the required computational effort;
- the average size of the human shoulders has been approximated to $hs = 60$ centimeters;



(a) (b)

Figure 4.4 (a) Virtual sensor designed for counting people crossing it in a given direction (identified by the arrow on the right). Activated cells due to the presence of the person are in blue. (b) The activation of the cells (in blue) and of the stripes (in red) in two consecutive frames.

- the number of stripes, as well as their size δ, has been automatically computed by analyzing the real size W_r (in centimeters) of the gate to be monitored (where the sensor is put). In more details, the number of cells is computed as follows:

$$K = \left\lceil \frac{L \cdot W_r}{hs} \right\rceil . \tag{4.4}$$

- H has been set so as to include in each cell the whole person. Note that a low H value would prevent the proposed approach to correctly work in cases of a low frame rate, since the risk is that the passage of a person is not detected in the two cells belonging to the same stripe. On the other side, a high H value would avoid to correctly detect the passage of two persons walking close to each other in a row. setting H as twice the depth of an average sized person represents a good tradeoff between the above two opposite requirements, as suggested in [108].

Thanks to these choices, the only parameters that the human operator has to set during the configuration step is the real width W_r of the gate, which can be easily obtained during the camera installation procedure. As evident, the lower is the number of parameters to set up, the higher is the possibility to use the system in real applications, where typically inexperienced human operators are required to install and configure such systems.

4.3 Gender Recognition

The module in charge of recognizing the gender can be partitioned in the following three steps: (1) face detection, (2) person tracking, (3) gender recognition. More details concerning each of the above steps is reported in the following, while the whole architecture is reported in Figure 4.3.

4.3.1 Face detection

Once a trigger from PC is received, the set of N images before the trigger and N following the trigger is processed and the faces are detected by the Viola Jones algorithm. This choice is due to the fact that the size of the face is expected to grow up after the trigger, while the person is walking towards the camera. In the experimental analysis, it will be demonstrated that the higher is the resolution of the face, the better the recognition performance is. Thus, the faces of the persons are selected as larger as possible (the images and thus the faces after the trigger). However, faces close to the camera may be much more affected by the distortion of the camera itself and by the motion blur, thus they could become more difficult to be detected by the Viola Jones algorithm. For this reason, the method keeps the images acquired both before and after the trigger.
In order to reduce the area of the image where searching for, and then the computational effort due to the detection step, the Viola Jones is not performed on the whole image I but instead only over a small region of the image, namely $A(I)$. $A(I)$ corresponds to the projection of the stripes in the sets AS which contribute to the counting of a person:

$$A(I) = proj_{AS}(I). \tag{4.5}$$

An example is shown in Figure 4.2. On the left, the head view image acquired by PC-cam, partitioned into five stripes (from S_1 to S_5), three of them in red since activated by the presence of the person, is reported. On the right, the image acquired by GR-cam,

together with the area of the scene corresponding to the activated area $A(i)$ (in red), is reported. Such area covers about one third of the whole image, thus implying a strong improvement in terms of performance.

4.3.2 Person Tracking

In order to identify the sequence of faces belonging to each person, a one-to-one overlapping tracking algorithm is applied. Tracking persons when dealing with videos instead than with still images is required for two main reasons: (1) the system which collects data extracted from video analytic systems should receive the gender of each person (and not the gender of each face) in order to extract useful statistics related for instance to the percentage of women entering a shopping center; (2) the decision concerning the gender of a person has to be taken by evaluating all the faces of that person and not just by evaluating a single face of the person.

In more details, a similarity matrix S is computed; the generic element $S(i, j)$ represents the similarity between the j-th face detected at the current frame and the i-th person detected until the previous frame. The similarity is based on the euclidean distance d between the position of the j-th detected face and the position of the face associated to i-th person at the previous frame:

$$S(i,j) = \begin{cases} 1 - \frac{d}{d_{max}} & \text{if } d \leq d_{max} \\ 0 & \text{otherwise} \end{cases} \tag{4.6}$$

being d_{max} the maximum displacement of a person between two consecutive frames. An example is shown in Figure 4.5.

Given S, the maximum value at the generic position (m, n) is computed and the association between the n-th detected face and m-th the person is performed, so that the information related to the person is updated with the new face. In cases a face can not be associated to any person, then a new identifier is associated to

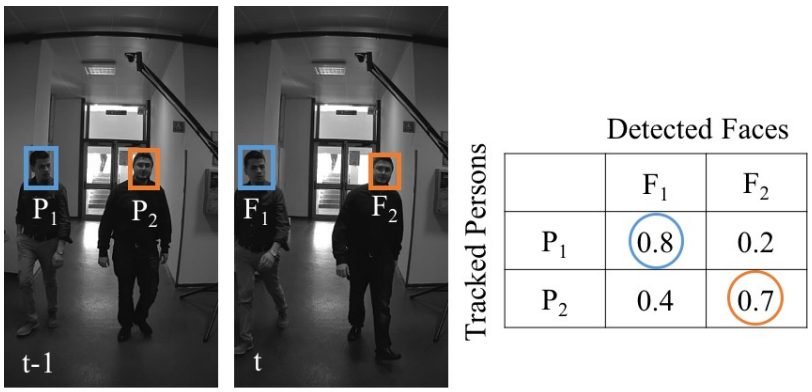


Figure 4.5 The similarity matrix is computed: the detected face F_1 is associated to the person P_1 (with a similarity value equal to 0.8), while the face F_2 is associated to the person P_2 (with a similarity value equal to 0.7).

that face and the algorithm starts tracking this new person.

Note that, differently from traditional object tracking, this problem is more simple since there are not any splits or merge that need to be solved [128]. Indeed, the only problem is related to the so called *ghosts*, namely to those objects (in our context those persons) which disappear for one or more frames due to some missing errors during the detection step.

In order to address this problem, the proposed algorithm does not stop tracking a person immediately (that is as soon as the association between a face and a person is performed) but only after t seconds, being t set in the experiments to 1 second.

4.3.3 Gender recognition from face images

The classification is based on a shape descriptor, namely the HOG descriptor. This choice can be considered a good trade-off between the accuracy achievable by this kind of descriptor and the computational burden due to its calculation.
In more details, the face is resized to 128×128 pixels and the HOG descriptor is extracted. Firstly the gradient magnitude and angle

of every pixel is computed by considering the responses of first-order partial derivatives of a 2D Gaussian function with a $\sigma = 1$. Then blocks of 32×32 pixels that overlap by 50% are computed and for each block a spatial tiling of 2×2 is used. For each tile the L2-normalized weighted histogram of 9 bins is computed (in intervals of 20 degrees), the normalized values clipped at 0.2 and normalized again. Considering that the size of the face images is 128×128 pixels, the HOG descriptor results in a (7 blocks $\times$ 7 blocks $\times$ 4 tiles $\times$ 9 bins =) 1764-element vector. The classification is then performed by an SVM with a RBF kernel (having $C = 2$ and $\gamma = 0.1$), which maximizes the performance in the problem at hand.

Finally, in order to increase the overall reliability of the proposed approach and to provide for each person (instead that for each face) its gender, the whole sequence of $|F|$ faces $F = \{f_1, ..., f_{|F|}\}$ associated to each person is evaluated. For each face f_i, a class $c_i \in \{Male, Female\}$ is computed. Given the set of $|F|$ classes $C = \{c_1, ..., c_{|F|}\}$ a majority vote classifier is exploited. The main idea lies in the fact that the generic class c_i corresponds to a vote; the class (Male or Female) which obtains the highest number of votes is the winner and is thus associated to that person.

4.4 Implementation optimizations

Nowadays several commercial chips are available directly integrated on the cameras: think, as an example, to Samsung Techwin, Axis, Hikvision or Texas Instruments, which make available to third part developers an SDK for integrating directly on board their video analysis applications. A common feature is that all of them are based on SoC architectures, composed by general purpose CPUs, DSP, Ethernet controllers, serial and parallel ports, USB ports, flash memory, SDRAM, I/O processors. In more details, (i) a floating point unit is not provided; (ii) the CPUs typically provide SIMD instructions, allowing to process for each cycle at

least four pixels.
In order to deal with the above constraints, as suggested in [101], SIMD instructions are exploited by taking advantage of RAPP [129], an ANSI C library optimized for running directly over SIMD based CPU. The RAPP library is meant to provide an optimized and reliable computational backend for low-level processing. The interface is designed to allow hardware-accelerated implementations, while still being simple enough for easy deployment from higher-level code. The main benefit deriving from this choice is that all the pixelwise operations which need nested cycles are substituted with SIMD instructions. As experimentally proved in [101], the average improvement in terms of processed pixels per second by using this type of instructions is around one order of magnitude.
The second optimization has the aim of avoiding floating point operations in our algorithm, by substituting them with full integer operations. In this case, as proved in [101], the average improvement achieved is up to two orders of magnitude.
As for the people counting algorithm, the advantage of these optimizations is that all the pre-processing (gaussian filter), background updating (pixel by pixel), foreground extraction and post-processing (morphological operators) are performed by using SIMD instructions and full integer operations, making the algorithm very fast. Moreover, the background is not updated every frame, but with a temporal decimation defined at configuration time. The gender recognition algorithm, instead, takes advantage of a fast computation of the integral image, an optimized implementation of the Viola-Jones algorithm and a very efficient coding of the HOG descriptor.

Chapter 5

Experimental Results

5.1 Datasets

For the experimental analysis of the proposed approaches five different datasets have been used, namely GENDER-FERET, LFW, UNISA-Public, UNISA-Private and SM-Private. GENDER-FERET and LFW have been used for evaluating the methods on still images. while UNISA-Public, UNISA-Private and SM-Private have been used for analyzing the performance in real-life scenarios. More details about the composition of the datasets are reported in the following subsections.

5.1.1 GENDER-FERET

In order to aim for some standardization, the GENDER-FERET subset, created from the well known FERET [20] dataset and already publicly available[1], has been used for the experimental analysis. The GENDER-FERET dataset has a balanced number of male and female face images and it is pre-partitioned into 474 training (237 males and 237 females) and 472 test (236 males and 236 females) images. This dataset consists of frontal faces acquired in controlled conditions with different illumination, background,

[1]The dataset is available under request at the following link: http://mivia.unisa.it

Figure 5.1 Examples of GENDER-FERET images

Figure 5.2 Number of training and test images, with details about male and female faces, used for experiments on the GENDER-FERET dataset.

Dataset	Training set	Test set	Total	
GENDER-FERET	237	236	473	**M**
	237	236	473	**F**
	474	472	946	**Total**

age, expression and race. Moreover, it contains only one face for each person, which can be present either in the training or in the test set, but not in both. Figure 5.1 shows some examples of face images from the GENDER-FERET dataset.

Table 5.2 reports the details of the GENDER-FERET dataset in terms of number of training and test, male and female face images that we used for our experiments.

5.1.2 LFW

In order to test the proposed method on faces with more different poses and in order to evaluate the impact of the face alignment technique, also the Labeled Faces in the Wild (LFW) dataset [21] has been considered for the experiments. LFW contains more than $13,000$ images of $5,749$ subjects designed to study the problem of face recognition in uncontrolled conditions. The images show famous people busy in different activities, such as recording an interview, playing sports, doing a fashion show and others. In the experiments the LFW face images have been aligned by using the algorithm described in Section 3.1.

Figure 5.3 Examples of LFW faces (first row) and the corresponding aligned images (second row)

Figure 5.4 Number of training and test images, with details about male and female faces, used for experiments on the LFW dataset.

Dataset	**Training set**	**Test set**	**Total**	
LFW	5976	1494	7470	**M**
	1834	459	2293	**F**
	7810	1953	9763	**Total**

Figure 5.3 shows four original LFW images and the corresponding aligned faces. The Viola-Jones algorithm has been applied to all the aligned LFW face images and the faces have been detected in 9,763 images. The groundtruth ha been generated by manually labelling these images as males and females, accordingly[2]. Then, for coherency with other methods like [45], [38] and [48], a 5-fold cross validation has been performed and the average accuracy of the proposed method has been computed.

Table 5.4 reports the details of the LFW dataset in terms of number of training and test, male and female face images used for our experiments.

[2]The dataset is available under request at the following link: http://mivia.unisa.it

5.1.3 UNISA-Public, UNISA-Private and SM-Private

The LFW images are unconstrained pictures and are taken with professional cameras. So the dataset consists of high resolution faces without motion blur, which is in contrast with real videos that are characterized by lower resolution and motion blur. For this reason, other types of unconstrained videos are needed for the evaluation of the algorithms in real scenarios.
Currently there are not any datasets available in the literature composed by videos (and not by still images) acquired in real environments. Indeed, most of the datasets are only composed by a set of images (of course more images for each person, but not necessarily extracted from consecutive frames). Often, these frames mainly contain a single face, which typically covers more than one half of the whole image, thus they are very different from a frame that could be instead acquired in a real environment. Furthermore, there are not any datasets that combine synchronized images acquired by two different cameras mounted in frontal view (for gender recognition) and head view (for people counting), respectively.

For this reason, a new dataset has been acquired; in more details, a Samsung SND-6084 (with a 3mm focal length) has been used as PC-cam and a pinhole Samsung SNB-6010 as GR-cam. Images from PC-cam have been acquired with a 1CIF resolution ($320x240$), at a frame rate of 25 frames per second. An higher resolution ($1080x1920$) has been used for acquiring images from GR-cam, but a lower frame rate (5 fps). This choice is justified by the fact that in case of gender recognition, the resolution has a high priority if compared with the frame rate. Furthermore, in order to reduce the motion blur in the acquisition of the faces, that typically makes worse the performance of gender recognition systems, the shutter time has been set to 5 ms.

The dataset has been acquired in three environments: (1) in two different gates of the University of Salerno, namely the Engi-

(a) UNISA-Public (b) UNISA-Private (c) SM-Private

Figure 5.5 The same person acquired in the UNISA-Public, UNISA-Private and SM-Private datasets

neering faculty and the Humanistic faculty, so as to have a balance between males (more frequent in the Engineering faculty) and and females (more frequent in the Humanistic faculty). In both the scenarios, the images have been acquired with different illumination conditions, in overexposed environments close to entrance or exiting doors and underexposed spaces like dark corridors. As expected in a context like a university, most of the persons are young people (in the range 18-30). We will refer hereinafter to this dataset as UNISA dataset. Note that, due to some privacy issues, only a part of this dataset can be made available [3]. Thus, in order to allow future benchmarking, the dataset has been partitioned into two different parts: *UNISA-public* refers to the part of the dataset that we have made publicly available, while *UNISA-private* refers to the remaining part. (2) The other environment where the dataset has been acquired is a supermarket in Salerno, where most of the people are adults or elders (in the range 40-70).

[3]The dataset is available under request at the following link: http://mivia.unisa.it

Figure 5.6 Description of the UNISA-Public, UNISA-Private and SM-Private datasets. *M-F* in the second and third columns refer to the number of males and females, respectively.

Name	# Persons (M-F)	# Faces (M-F)
UNISA-public	$60(43-17)$	$430(300-130)$
UNISA-private	$693(367-326)$	$3775(1797-1978)$
SM-private	$147(71-76)$	$1627(751-876)$

Note that, due to some privacy issues, even this dataset can not be made publicly available, and it is called *SM-private*. More details concerning the datasets are provided in Table 5.6, while an example for each dataset is provided in Figure 5.5. In the whole, more than 5600 faces have been acquired from about 900 different persons. The dataset is very challenging, since the faces exhibit all the possible variations of a face captured by a camera. People often do not look at the camera and their faces are detected with different poses and expressions. For the same reason, the persons make sudden movements that cause motion blur on the image, or involuntarily occlude their faces with hands, handkerchiefs and smartphones. Furthermore, people are recorded at different distances from the camera (ranging from half a meter to 4 meters), so the face images have different resolutions.

5.2 Experiments on still images

In this section, the effectiveness of the gender recognition methods on the GENDER-FERET and the LFW datasets is evaluated. First, the results of the fusion of handcrafted features are reported. Then, the results of the combination of trainable COSFIRE filters and the SURF-based methods are reported. Finally, the results are compared between them and with other scientific methods and commercial libraries.

Figure 5.7 Experimental results with a fusion of handcrafted features. The last two columns report the accuracy rate achieved on the two datasets.

Raw	LBP	HOG	GENDER-FERET	LFW
✓			89.6	96.6
	✓		89.6	93.6
		✓	90.0	97.2
✓	✓		91.9	97.5
✓		✓	92.6	98.4
	✓	✓	91.5	97.1
✓	✓	✓	**93.0**	**98.4**

5.2.1 Results with a fusion of handcrafted features

As shown in Table 5.7, the accuracy of the methods based on handcrafted features is above 89% using any combination of feature or classifier. On the GENDER-FERET dataset the classifiers that relies on only raw pixel values and LBP histograms achieve 89.6% of accuracy, which is a little bit worse than the HOG-based classifier, that achieves 90.0% of accuracy. Performance significantly increases when the decisions of different classifiers are combined. The best performance, equal to 93.0%, is achieved combining all the classifiers.
The same trend can be observed by evaluating the results on the LFW dataset. The single expert with the best performance is still the HOG-based classifier, that achieves 97.2% of accuracy. On this dataset the LBP-based expert is not very effective like the other two and this result is evident even evaluating the performance of the MESs. Indeed, the combination of the raw- and the HOG-based classifiers (98.4%) outperforms the two multi-experts which rely on the LBP classifier (97.5% and 97.1%) and achieve the same accuracy of the multi-expert which combine all the decisions.

5.2.2 Choosing the optimum number of COSFIRE filters

For the experiments with the COSFIRE-based method, the GENDER-FERET has been preliminarily used to find the optimum number of COSFIRE filters needed to describe the face. First, five training faces of men and five training faces of women have been chosen. Then, for each randomly picked face a random region of size 19×19 pixels has been chosen and used as a prototype to configure a COSFIRE filter. If the selected prototype resulted in a COSFIRE with less than 5 tuples, it was considered as not enough salient and a new one was chosen. The filters were configured with the default parameters $t_1 = 0.1$, $t_2 = 0.75$, $\sigma_0 = 0.67$ and $\alpha = 0.1$ as proposed in [118]. In the configuration of the filters were considered Gabor filter responses along three concentric circles and the center point: $\rho = \{0, 3, 6, 9\}$. The sizes of the prototype patterns together with the number and radii of the concentric circles were determined empirically on the training set. Then further experiments have been executed by incrementing the set of COSFIRE filters by 10 at a time up to 250. In Figure 5.8 the accuracy rate as a function of the number of filters used is plotted. For each set of COSFIRE filters two values are depicted, one of which is the training accuracy rate that is achieved by 10-fold cross validation on the training set, and the other one is the accuracy rate obtained on the test set. With only 10 filters that result in a feature vector of $(21 \times 10 =)$ 210 elements we achieved 83.79% and 81.4% accuracy rates on the training and test sets, respectively. The accuracy increased rapidly up to 60 filters and then increased slowly until it reached a plateau. The maximum accuracy rate was achieved with 180 COSFIRE filters. By using the same 180 filters 94.1% of accuracy was achieved on the test set.

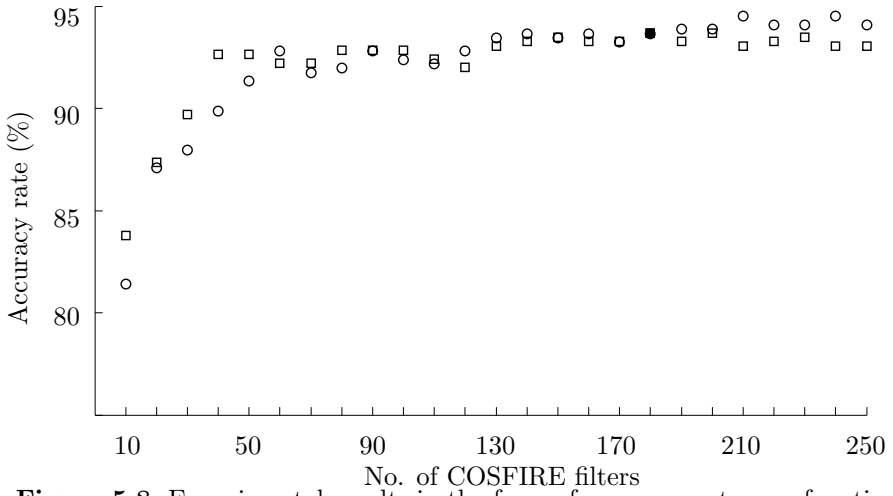


Figure 5.8 Experimental results in the form of accuracy rate as a function of the number of COSFIRE filters used. The square markers indicate the accuracy rate on the training set with a 10-fold cross validation while the circles indicate the accuracy rates on the test set. The solid square marker indicates the maximum accuracy rate on the training set, which is achieved with 180 filters.

5.2.3 Results with the fusion of COSFIRE- and SURF- based classifiers

As described above, for the experiments with trainable features 90 COSFIRE filters from randomly selected male face training images and 90 from randomly selected female face training images have been configured both for the GENDER-FERET and the LFW datasets. The SURF-based algorithm, instead, does not require the configuration of any parameters.

Table 5.9 shows the results of the COSFIRE-based and of the SURF-based methods on the GENDER-FERET and the LFW benchmark datasets. The first one consistently outperforms the domain-specific method that uses the SURF descriptor for the fiducial landmarks. In particular, it achieves an accuracy of 94.1% on the GENDER-FERET dataset and a remarkable accuracy of 99.3% on the LFW dataset. Indeed the SURF-based approach achieves a high accuracy rate on the LFW dataset (96.1%), but a

Figure 5.9 Results of the COSFIRE- and SURF-based methods on the GENDER-FERET and the LFW datasets.

Dataset	Method	Accuracy (%)
GENDER-FERET	COSFIRE-based	94.1
	SURF-based	89.2
	COSFIRE+SURF	**94.7**
LFW	COSFIRE-based	99.3
	SURF-based	96.1
	COSFIRE+SURF	**99.4**

Figure 5.10 Accuracy achievable by fusing COSFIRE- and SURF-based classifiers.

moderate performance on the GENDER-FERET (89.2%). In both the cases the stacked classification scheme, which combines the decisions of the two experts, is able to improve the accuracy of the single experts. So the best performance of the proposed method is 94.7% on the GENDER-FERET and 99.4% on the LFW.

5.2.4 Evaluating the complementarity of COSFIRE- and SURF-based classifiers

In this section the complementarity of the trainable COSFIRE- and the domain-specific SURF-based approaches is analysed. Indeed, the potential of the multi-expert has to be firstly evaluated, since in this case the features are not clearly complementary like the raw, texture and shape features. In particular, Figure 5.10

gives an idea of the performance achievable by combining the decisions of the proposed classifiers over the two considered datasets. The ideal fusion technique would be the one that is able to take the right decision when at least one of the methods classifies the gender correctly. This happens in two situations: (1) both the decisions are correct (in blue in the figure); (2) only one of the two decisions is correct (in red in the figure). Such experiment proves the complementarity of the COSFIRE- and the SURF-based features, since the fusion may achieve more than 99% of accuracy both on the GENDER-FERET and the LFW datasets.

5.2.5 Evaluating the effectiveness of the combination rules

In order to justify the choice of the weighted voting rule for the fusion of handcrafted features and the stacked classification scheme for the combination of trainable and domain-specific classifiers, the performance of the different combination techniques have been evaluated on the two considered datasets.
For each face image, the application of the majority voting rule results in the sum of the male and female probabilities and in the evaluation of the total probabilities. If the total male probability is greater than the total female probability, then the given face image is labelled to be a male otherwise a female. The probability of a decision taken by a classifier is computed with a sigmoid function, by using the SVM score returned by the classifier (increasing the distance from the margin, the probability increases and other way around). For the weighted voting rule, such probabilities are multiplied with the prior probabilities (namely the accuracies) on the training sets achieved by the single classifiers on the male and female classes and then summed. The classifier which yields the highest absolute weighted score is entrusted. As for the stacked classification scheme, a linear SVM with $C = 1$ has been used to learn the final classification level from the decisions taken by the classifiers on the training set.
Table 5.11 reports the results achieved on the two datasets using

Figure 5.11 Results with three fusion techniques on the GENDER-FERET and the LFW datasets. In the second column, MV indicates the majority voting, WV the weighted voting and SC the stacked classification rules. In the third and in the fourth columns, respectively, RLH represents the accuracy rate achieved by the fusion of raw-, LBP- and HOG-based classifiers, while CS indicates the results obtained by the combination of COSFIRE- and SURF-based experts.

Dataset	Rule	RLH	CS
GENDER-FERET	MV	93.0	94.5
	WV	**93.0**	94.3
	SC	92.6	**94.7**
LFW	MV	98.3	99.0
	WV	**98.4**	99.0
	SC	97.9	**99.4**

the three different fusion methods. The experiment points out that the fusion of handcrafted features is able to achieve the best performance when the decisions are combined with the weighted voting rule. Substantially, both the majority and the weighted voting achieve an higher accuracy than the stacked classification with these sets of features. On the other hand, the fusion of trainable and domain-specific classifiers is more effective by using the stacked classification scheme, even if it is not able to achieve the ideal performance depicted in Figure 5.10. In this case, the majority voting and the weighted voting schemes obtain almost the same performance and, even, achieve lower accuracies results than that of the COSFIRE-based method on the LFW dataset.

5.2.6 Comparison with other methods

In order to prove the effectiveness of the proposed approaches with respect to other methodologies, a comparative analysis has been performed. Table 5.12 shows the performance comparison on the GENDER-FERET dataset. The comparison has been performed between the proposed approaches and two commercial libraries, namely Face++ [130] and Luxand [131]. Face++ is a well-known

Figure 5.12 Comparison of the results on the GENDER-FERET dataset.

Method	**Description**	**Accuracy (%)**
RAW	Intensity	89.6
LBP	Texture	89.6
HOG	Shape	90.0
RAW LBP	Intensity Texture	91.9
RAW HOG	Intensity Shape	92.6
LBP HOG	Texture Shape	91.5
RAW LBP HOG	Handcrafted fusion	93.0
COSFIRE	Trainable shape	94.1
SURF	Facial landmarks	89.2
COSFIRE SURF	Trainable and handcrafted	**94.7**
Face++	[130]	89.6
Luxand	[131]	89.2

and widely adopted library based on a deep learning approach. On the other side, Luxand is based on geometric features: indeed, 70 facial features which describe the position of salient points of the face are used.
The trainable COSFIRE-based approach outperforms the method which exploits the use of pixel intensity values, texture and shape features. Moreover, the fusion with the SURF-based method further improves the results. Most of the proposed descriptors are able to outperform the commercial libraries, especially the HOG- and the COSFIRE-based classifiers. This result suggests that the shape features are able to better discriminate between men and women. However, the best result is achieved by the combination of trainable COSFIRE filters and SURF descriptors.

In Table 5.13 the results have been compared with existing approaches on the LFW dataset. Most of the proposed methods are more effective than the approaches proposed by Dago-Casas et al. [45] and Shan et al. [38], while only the fusion of handcrafted features and trainable and SURF features outperform the approach proposed by Tapia and Perez [48]. It is worth to note that the best performance is achieved by the combination of trainable and

Figure 5.13 Comparison of the results on the LFW dataset.

Method	Description	Accuracy (%)
RAW	Intensity	96.6
LBP	Texture	93.6
HOG	Shape	97.2
RAW LBP	Intensity Texture	97.5
RAW HOG	Intensity Shape	98.4
LBP HOG	Texture Shape	97.1
RAW LBP HOG	Handcrafted fusion	98.4
COSFIRE	Trainable shape	99.3
SURF	Facial landmarks	96.1
COSFIRE SURF	Trainable and handcrafted	**99.4**
Dago-Casas et al. [45]	Gabor	94.0
Shan et al. [38]	Boosted LBP	94.8
Tapia and Perez [48]	LBP	98.0

handcrafted features extracted from facial landmarks, even if the most important contribution is provided by the COSFIRE-based classifier.
As shown in the table, all the methods use SVM classifiers. Thus, the main improvement is due to the combination of the chosen descriptors, which proved to be very representative for the problem at hand.

5.3 Experiments on images extracted from real video sequences

In this Section the results obtained by the proposed architecture on images extracted from real video sequences are reported. The tests have been conducted by using two SoC platforms. As for the gender recognition module the ARMv8 Cortex-A53 1.2 GHz equipped by the Raspberry Pi3 has been adopted. On the other hand, the people counting module has been evaluated by installing the application directly on board of the Samsung Wisenet III DSP provided by the SND-6084 camera used for the acquisition.
In order to confirm the effectiveness of the proposed architecture

Figure 5.14 Experiment 1: accuracy obtained for both single faces and persons over the images of UNISA-Public dataset triggered by the people counting algorithm. The training has been performed by using the UNISA-Private and the SM-Private datasets. M, F and $Tot.$ refer to the accuracy for Male, Female and to the overall accuracy, respectively.

	Acc. per face (%)	Acc. per person (%)
M	92.3	93.0
F	78.5	76.5
Tot.	84.5	88.3

as well as the reliability of the proposed gender recognition module, various experiments have been performed. Note that for each experiment the results achieved by analyzing both the *faces* and the *persons* are reported. In the case of accuracy per face, only spatial information are exploited. It means that each face is analyzed independently on the other belonging to the same person and the obtained accuracy, normalized with respect to the number of faces in each dataset, is reported. On the other hand, the accuracy per person is computed by evaluating spatio-temporal information, including the tracking step and the majority voting classifier. In this way, for each person only a single vote (male or female) is given, and not one vote for each frame in which the person is framed. Of course, the accuracy in this case is normalized with respect to the number of persons and not to the number of faces.

5.3.1 Generalization capabilities

In the first experiment the proposed approach has been evaluated over the UNISA-Public dataset, while face images from both UNISA-Private and SM-Private have been used for training, so as to verify the capability of the proposed approach to generalize in different scenarios. In more details, only the set of images triggered by the people counting algorithm have been considered (thus by considering 11 images per person, namely 5 before the trigger, 1 in the time instant of the trigger and 5 after the trigger). In this way, the proposed architecture has been tested in terms of

reliability. To this concern, the evaluation took into account: (i) the accuracy of the considered people counting algorithm; (ii) the accuracy in the recognition of the faces; (iii) the accuracy in the recognition of the gender of the persons.
(i) As for the first point, 100% of the passages has been correctly found, without any miss, confirming the effectiveness of the adopted people counting algorithm even if used over an embedded architecture.
(ii) Second, the faces have been found in 95% of the cases, using the Viola-Jones algorithm fully implemented with RAPP library (by porting over RAPP the OpenCV implementation), so as to obtain the maximum optimization over embedded architecture. Viola-Jones has been configured so as to search faces with a resolution in the range 80×80 and 220×220 pixels. In order to reduce the number of false positives, the so called *min neighbors* parameter, corresponding to the neighbors that each candidate rectangle should have in order to retain it, has been set to 7 and the *scaling factor* to 1.1. However, considering that our evaluation is not based on a single face but instead on the whole sequence of faces associated to each person, the results have been analysed also in terms of recognition of the person, which in turn correspond to recognize each person for at least one frame. In the experiments, 100% of the persons has been correctly detected for at least one frame.
(iii) Finally the performance over all the detected faces (and persons) has been evaluated. The accuracy is reported both in terms of faces and persons, by considering a majority voting approach for all the faces found for that person.
The results are reported in Table 5.14. The accuracy in the recognition of males is higher than the ones of females (92.3% vs 78.5%). It is mainly due to the lower number of images available in the test set for females (130 for females vs 300 for males). However the overall accuracy obtained by the proposed architecture is very promising (88.3%) considering how challenging the dataset is.

Figure 5.15 Experiment 2: accuracy obtained for both single faces and persons over the videos available in the UNISA-Public dataset.

	Acc. per face (%)	Acc. per person (%)
M	88.4	88.1
F	77.5	82.3
Tot.	85.6	86.4

5.3.2 Experiments on the UNISA-Public video sequences

In order to confirm the usability of the proposed architecture on video sequences and to give the possibility to other researchers to compare the results, the gender recognition algorithm has been evaluated over the whole set of images of the UNISA-Public dataset (instead than just over the 11 images triggered by the people counting algorithm). The accuracy achieved is reported in Table 5.15. It is important to highlight that by testing over the whole video instead than over a few images the *ideal* condition is simulated, but not the *real* situation. Indeed, the frame rate reached over the considered SoC architecture with a full resolution is less than 1 frame per seconds, which is lower than the 5 frames per second of the video. In order to verify the possibility to further optimize the proposed algorithm, the time required by the different steps of the gender recognition algorithm is also analysed. However about 85% of the time is spent during the detection step (by the Viola-Jones algorithm). It implies that the time can not be further optimized, thus justifying the introduction of the proposed multi-sensor architecture.

5.3.3 Impact of resolution on the accuracy

In order to confirm that a high resolution of the faces is required to achieve a high accuracy in both face detection and gender recognition, an evaluation about how the accuracy scales with respect to the resolution of the images is reported. In particular, three resolutions have been considered: *1080x1920* (which is the full resolution of the video), *480x640* (4CIF) and *240x320* (1CIF).

Figure 5.16 Experiment 3: accuracy obtained for both single faces and persons by varying the resolution of the images. *Det.* refers to the percentage of faces detected using the Viola Jones algorithm.

	Acc. per face (%)			Acc. per person (%)		
	FULL	4CIF	1CIF	FULL	4CIF	1CIF
Det.	72.2	71.7	24.9	95.0	91.7	63.3
M	88.4	85.6	75.0	88.1	83.3	52.4
F	77.5	82.1	87.5	82.3	76.5	41.2
Tot.	85.6	84.7	78.1	86.4	81.4	49.1

The Viola-Jones algorithm has been configured as in the first experiment, scaling the range of the face size by a factor 2 on the 4CIF images and 4 on the 1CIF images. The obtained results, reported in terms of accuracy computed by evaluating both single faces and persons, are shown in Table 5.16. The first important result (first row of the table, *Det.*) is related to the percentage of detected faces, which strongly drops by decreasing the resolution (from 72.2% with a full resolution to 24.9% with 1CIF images). It implies that with low resolution images (1CIF), about 40% of the persons is completely missed, while with full resolution images only 5% of the persons is not detected.
The importance in processing high resolutions images is also confirmed by analyzing the results in terms of gender recognition rate. The gender of less than one half of the persons is correctly recognized (49.1% of accuracy) by using 1CIF images, while more than 86% by analyzing the face extracted by full resolution images. This very impressive result confirms the importance of analyzing high resolution images.

5.3.4 Experiments on the whole dataset

The tests performed up to now are also based on the UNISA-Public dataset, which is publicly available. It allows to compare any other algorithm with the proposed approach. However, in order to confirm the effectiveness of the proposed approach, it has been also tested with the two private datasets, namely UNISA-Private and SM-Private. Each dataset has been partitioned in training

Figure 5.17 Experiment 4: Accuracy for face and for person obtained over UNISA-Public (UPub), UNISA-Private (UPri) and SM-Private (SPriv).

	Acc. per face (%)			Acc. per person (%)		
	UPub	UPriv	SPriv	UPub	UPriv	SPriv
M	93.2	91.9	87.3	95.2	92.0	80.3
F	98.4	95.0	78.1	100.0	95.8	82.9
Tot.	94.7	93.0	82.4	96.5	93.9	81.6
Overall	90.2			90.5		

Figure 5.18 Comparison of the proposed approach with state of the art.

Method	Reference	Accuracy (%)
Face++	[130]	75.4
Luxand	[131]	77.1
Beta Face	[132]	77.8
Kairos	[133]	81.7
Microsoft Face API	[134]	85.7
Proposed method	-	**90.2**

(50%) and test (50%) set, by avoiding that the same person is both in the training and in the test set, even if in different places or in different poses and illumination conditions. The achieved results are reported in Table 5.17. The overall accuracy in the three datasets is higher than 80% (from 81.6 to 96.5), and in the whole it is equal to 90.5%, so confirming the effectiveness and thus the applicability of the proposed approach in different scenarios.

5.3.5 Comparison with other methods

In order to further confirm the effectiveness of the proposed approach, a comparative analysis with available commercial libraries (such as Face++, Luxand, BetaFace, Kairos, Microsoft Face API) has been performed. For Kairos, Microsoft Face API and Beta Face the details about the algorithms are not available, while Face++ and Luxand have been described above.
Table 5.18 shows that the proposed method is able to outperform both these approaches of more than 10%. In general, the proposed approach overcomes the best method of more than 5%, confirming its effectiveness and usability in real scenarios.

Figure 5.19 Processing time and frame rate of RAW-, LBP-, and HOG-based classifiers on 640 × 480 video sequences executed on an Intel(R) Core(TM) i7-3770S CPU @ 3.10 GHz 4GB RAM.

	RAW (ms)	**LBP (ms)**	**HOG (ms)**
Loading image	8.5		
Detection	53.1		
Alignment	177.4		
Resize	0.2		
Description	2.4	16.1	4.2
Classification	12.6	11.1	8.7
Total (ms)	254.2	266.4	252.1

5.4 Evaluating the processing time

In this section an extensive analysis of the processing time required by some of the proposed classifiers is provided. Such evaluation explains the rationale behind the choice of the multi-sensor architecture.

Table 5.19 reports the computational requirement of the RAW-, LBP- and the HOG-based classifiers on an Intel(R) Core(TM) i7-3770S CPU @ 3.10 GHz 4GB RAM with 640×480 video sequences. The results point out that the most costly processing steps are the face detection and alignment. The latter requires about 177 milliseconds, more than three times the amount required by the detection, which is an essential step. For this reason, in the method proposed for the analysis of video sequences the face alignment step has been removed.

The Table also allows to compare the performance of the different face descriptors. The raw features ($2.4ms$), obviously, are faster than the LBP ($16.1ms$) and the HOG ($4.2ms$) ones. However, they requires more time in the classification step (12.6 vs 11.6 and 8.7 milliseconds) since the descriptor has an higher dimensionality (4096 vs 2304 and 1764 elements). In general, the HOG-based classifier is the fastest and it has been chosen for gender recognition on video sequences. The SURF-and the COSFIRE-based classifiers have not been considered in this analysis, since they require, respectively, 150 milliseconds and more than one second for the computation of a single face descriptor.

Figure 5.20 Processing time and frame rate with and without face alignment of the HOG-based classifier on video sequences at different resolutions executed on an Intel(R) Core(TM) i7-3770S CPU @ 3.10 GHz 4GB RAM.

	1920 × 1080		640 × 480		320 × 240	
HOG	ms	fps	ms	fps	ms	fps
With alignment	425.5	2.3	252.1	3.96	203.1	4.9
Without alignment	214.6	4.7	74.7	13.4	36.9	27.1

Figure 5.21 Processing frame rate of people counting and gender recognition algorithms on video sequences at different resolutions executed respectively on a ARMv7 with 1CPU 600 MHz 0.8GB RAM and on an ARMv8 Cortex-A53 1.2 GHz.

	1920 × 1080		640 × 480		320 × 240	
Multi-sensor	ms	fps	ms	fps	ms	fps
People counting	256.4	3.9	61.0	16.4	15.5	64.5
Gender recognition	2017.2	0.5	709.6	1.4	348.7	2.9

Table 5.20 shows the improvement of the performance in terms of processing speed removing the face alignment step. On 640×480 the frame rate increases from about 4 fps to more than 13, so allowing almost to carry out real-time gender recognition. The ideal real-time speed is achieved on 320×240 images, since the algorithm is able to analyze about 27 fps. Also the performance on 1920×1080 is notable, considering that the proposed method processes almost 5 high definition frames per second. So the approach is definitely suitable for server side elaboration and, in principle, it would not need the help of the people counting algorithm. However, the computation capabilities of an embedded system are in general lower than a powerful server devoted to video analysis.

Table 5.21 reports the processing time required by the people counting and the gender recognition algorithms on the two considered platforms, namely the ARMv7 with 1CPU 600 MHz 0.8GB RAM equipped by the smart camera and the ARMv8 Cortex-A53 1.2 GHz processor of the Raspberry Pi3. As published in [108], the proposed people counting is able to process 320×240 images in real-time, without paying a lot in terms of accuracy. The gender

recognition method, instead, points out a significant performance drop on such low cost embedded platform. This result means that real-time face analysis is not achievable on the considered device, since most of the time is spent for face detection (more than 70%), which can not be removed or further optimized without the help of other sensors.
Such experimental evaluation justifies the choice of the multi-sensors architecture. Indeed, the people counting algorithm allows to reduce the region where the faces are detected, so increasing the processing speed of this step, and to enable the elaboration only when a person is going towards the camera. The two optimizations makes the gender recognition algorithm capable to collect and process more images for each person, to analyze, if necessary, high resolution faces and to reduce false positives in face detection.

Chapter 6

Conclusions

6.1 Summary of the thesis

This research work has been produced with the aim of performing gender recognition in real-time on face images extracted from real video sequences. The task may appear easy for a human, but it is not so simple for a computer vision algorithm. Even on still images, the gender recognition classifiers have to deal with challenging problems mainly due to the possible face variations, in terms of age, ethnicity, pose, scale, occlusions and so on. Additional challenges have to be taken into account when the face analysis is performed on images acquired in real scenarios with traditional surveillance cameras. Indeed, the people are unaware of the presence of the camera and their sudden movements, together with the low quality of the images, further stress the noise on the faces, which are affected by motion blur, different orientations and various scales. Moreover, the need of providing a single classification of a person (and not for each face image) in real-time imposes to design a fast gender recognition algorithm, able to track a person in different frames and to give the information about the gender quickly. The real-time constraint acquires even more relevance considering that one of the goals of this research work is to design an algorithm suitable for an embedded vision architecture. Such necessity excludes the use of CNNs, since they

require computational and memory resources that are not available on smart cameras and low cost embedded devices. Finally, the task becomes even more challenging since there are not standard benchmarks and protocols for the evaluation of gender recognition algorithms.

In this thesis the attention has been firstly concentrated on the analysis of still images, in order to understand which are the most effective features for gender recognition. To this aim, a face alignment algorithm has been applied to the face images so as to normalize the pose and optimize the performance of the subsequent processing steps. Then two methods have been proposed for gender recognition on still images.

First, a multi-expert which combines the decisions of classifiers fed with handcrafted features has been evaluated. The pixel intensity values of face images, namely the raw features, the LBP histograms and the HOG features have been used to train three experts which takes their decision by taking into account, respectively, the information about color, texture and shape of a human face. The decisions of the single linear SVMs have been combined with a weighted voting rule, which demonstrated to be the most effective for the problem at hand.

Second, a SVM classifier with a chi-squared kernel based on trainable COSFIRE filters has been fused with an expert which rely on SURF features extracted in correspondence of certain facial landmarks. The complementarity of the two experts has been demonstrated and the decisions have been combined with a stacked classification scheme.

An experimental evaluation of all the methods has been carried out on the GENDER-FERET and the LFW datasets with a standard protocol, so allowing the possibility to perform a fair comparison of the results. Such evaluation proved that the couple COSFIRE-SURF is the one which achieves the best accuracy in all the cases, even compared with other state of the art methods. Anyway, the performance achieved by the multi-expert which rely on the fusion of RAW, LBP and HOG classifiers can also be considered very satisfying.

After the preliminary analysis carried out on still images, the research has been focused on video sequences. A new dataset, namely the UNISA-dataset, has been acquired in different real environments (university and supermarket) with classic surveillance cameras and a part of it has been made publicly available. In these video sequences people are unaware to be framed, so the face images are significantly more challenging than the ones available in the standard datasets.
Such benchmark has been used firstly for an extensive analysis of the processing time required by the above-mentioned gender recognition algorithms. The profiling activity demonstrated that the face alignment algorithm is very costly and it is not suitable for real-time elaboration, as well as the impossibility to use SURF and COSFIRE features. Considering that the pixel intensity values are not reliable with face images not aligned, the analysis also allowed to choose the HOG expert for gender recognition on video sequences, since it is more efficient and effective than the LBP one. Finally, the analysis shows that, although the HOG-based classifier is able to process images in real-time on classic server side architectures, it is not fast enough for embedded vision systems.
Starting from these observations, a multi-sensor architecture has been proposed. It consists of a smart camera dedicated to people counting, of a classic camera installed to capture the faces and of a low cost embedded device used for gender recognition. The idea is that the people counting camera sends a notification when a passage of at least one person is detected, indicating the position where the passage occurred. In this way, the gender recognition algorithm can be applied only on a subregion of the images where at least one face is present. Such architecture allows to process more high resolution images, so obtaining in parallel the possibility to maximize the accuracy and to recognize the gender of the persons in real-time on low cost devices. Moreover, the temporal coherency has been taken into account to associate the same identity to a person captured in different frames. An extensive experimental evaluation proved the effectiveness of the proposed

method and its suitability for gender recognition in real-time.

6.2 Future works

Although the proposed multi-sensor architecture demonstrated its effectiveness and efficiency, it would be better to achieve the same results with a single device or, at least, with a single camera. Such constraint would suggest to perform further optimizations on the gender recognition algorithm. An interesting idea to investigate would be the use of a background subtraction algorithm, which is able to detect pixels belonging to moving objects with fixed cameras. In this way, the notification about the position of the persons who are moving towards the camera is provided by the foreground mask and not by the people counting camera. The face detection can be consequently applied only on moving regions, so as to reduce the computational burden. However, extensive analysis have to be performed in order to check if the optimization is sufficient to process images in real-time on low-cost devices, preserving the performance in terms of accuracy.
So far as the optimization of the algorithm does not allow to achieve the desired result on low cost devices, an alternative would be the search of more powerful embedded systems. Technology in this field is growing exponentially and the market currently offers various solutions which permits to use several GPU cores for embedded vision applications. The most famous examples are Intel Movidius Neural Compute Stick and the NVIDIA Jetson TK1. Such powerful devices would allow to evaluate the possibility to use deep networks for gender recognition in real-time, which are obviously not suitable for the elaboration on classic CPUs and, above all, on low cost devices. The CNNs have not been considered in this thesis, due to their significant computational burden. Nevertheless, it is worth to carry out an experimental evaluation of deep networks on these modern devices, which may allow to achieve remarkable performance both in terms of accuracy and processing speed.

Finally, the research may be extended to other face analysis problem. For example, the same face descriptors, or other that are considered promising, can be applied to problems like age estimation, ethnicity classification, expression and sentiment analysis, face recognition, re-identification and verification. The experimental analysis proved that most of the time is spent for face detection. So on powerful platforms the computation of efficient descriptors or the classification of different facial features may not add a significant computational load.

Bibliography

[1] M. Ng, V. M. Ciaramitaro, S. Anstis, G. M. Boynton, and I. Fine, "Selectivity for the configural cues that identify the gender, ethnicity, and identity of faces in human cortex," *Proceedings of the National Academy of Sciences*, vol. 103, no. 51, pp. 19 552–19 557, 2006.

[2] S. Fu, H. He, and Z.-G. Hou, "Learning race from face: a survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 12, pp. 2483–2509, 2014.

[3] Y. Fu, G. Guo, and T. S. Huang, "Age synthesis and estimation via faces: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 11, pp. 1955–1976, 2010.

[4] M. E. Hasselmo, E. T. Rolls, and G. C. Baylis, "The role of expression and identity in the face-selective responses of neurons in the temporal visual cortex of the monkey," *Behavioural brain research*, vol. 32, no. 3, pp. 203–218, 1989.

[5] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern recognition*, vol. 36, no. 1, pp. 259–275, 2003.

[6] E. T. Rolls, M. J. Tovee, D. G. Purcell, A. L. Stewart, and P. Azzopardi, "The responses of neurons in the temporal cortex of primates, and face identification and detection," *Experimental Brain Research*, vol. 101, no. 3, pp. 473–484, 1994.

[7] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM computing surveys (CSUR)*, vol. 35, no. 4, pp. 399–458, 2003.

[8] C. A. Nelson, "The development and neural bases of face recognition," *Infant and child development*, vol. 10, no. 1-2, pp. 3–18, 2001.

[9] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[10] J. V. Haxby, E. A. Hoffman, and M. I. Gobbini, "The distributed human neural system for face perception," *Trends in cognitive sciences*, vol. 4, no. 6, pp. 223–233, 2000.

[11] C. B. Ng, Y. H. Tay, and B.-M. Goi, *Recognizing Human Gender in Computer Vision: A Survey*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 335–346.

[12] V. Bruce, A. Burton, E. Hanna, P. Healey, O. Mason, A. Coombes, R. Fright, and A. Linney, "Sex discrimination: how do we tell the difference between male and female faces?" *Perception*, vol. 22, no. 2, pp. 131–52, 1993.

[13] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 4, pp. 607–626, 2009.

[14] C. BenAbdelkader and P. Griffin, "A local region-based approach to gender classification from face images," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops - Volume 03*, ser. CVPR '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 52–. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2005.388

[15] U. Park and A. K. Jain, "Face matching and retrieval using soft biometrics," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 3, pp. 406–415, 2010.

[16] B.-C. Chen, C.-S. Chen, and W. H. Hsu, "Cross-age reference coding for age-invariant face recognition and retrieval," in *European Conference on Computer Vision*. Springer, 2014, pp. 768–783.

[17] D. Riccio, G. Tortora, M. De Marsico, and H. Wechsler, "Ega?ethnicity, gender and age, a pre-annotated face database," in *Biometric Measurements and Systems for Security and Medical Applications (BIOMS), 2012 IEEE Workshop on*. IEEE, 2012, pp. 1–8.

[18] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2170–2179, 2014.

[19] R. Azarmehr, R. Laganiere, W.-S. Lee, C. Xu, and D. Laroche, "Real-time embedded age and gender classification in unconstrained video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 57–65.

[20] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 10, pp. 1090–1104, 2000.

[21] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Technical Report 07-49, University of Massachusetts, Amherst, Tech. Rep., 2007.

[22] K. Ricanek and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," in *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*. IEEE, 2006, pp. 341–345.

[23] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 529–534.

[24] T. Sim, S. Baker, and M. Bsat, "The cmu pose, illumination, and expression (pie) database," in *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on.* IEEE, 2002, pp. 53–58.

[25] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

[26] A. M. Martinez, "The ar face database," *CVC technical report*, 1998.

[27] A. Kasinski, A. Florek, and A. Schmidt, "The put face database," *Image Processing and Communications*, vol. 13, no. 3-4, pp. 59–64, 2008.

[28] C. E. Thomaz and G. A. Giraldi, "A new ranking method for principal components analysis and its application to face image analysis," *Image and Vision Computing*, vol. 28, no. 6, pp. 902–913, 2010.

[29] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell, "Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on.* IEEE, 2011, pp. 74–81.

[30] M. Grgic, K. Delac, and S. Grgic, "Scface–surveillance cameras face database," *Multimedia tools and applications*, vol. 51, no. 3, pp. 863–879, 2011.

[31] S. Rudrani and S. Das, "Face recognition on low quality surveillance images, by compensating degradation," *Image Analysis and Recognition*, pp. 212–221, 2011.

[32] MiviaLab, "Gender-feret dataset," Available: http://mivia.unisa.it/database/gender-feret.zip, 2016.

[33] B. Moghaddam and M. Yang, "Learning gender with support faces," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 5, pp. 707–711, 2002.

[34] J. Yang, D. Zhang, A. F. Frangi, and J. Y. Yang, "Two-dimensional pca: a new approach to appearance-based face representation and recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 1, pp. 131–137, 2004.

[35] S. Baluja and H. A. Rowley, "Boosting sex identification performance," *International Journal of computer vision*, vol. 71, no. 1, pp. 111–119, 2007.

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

[37] Z. Yang and H. Ai, "Demographic classification with local binary patterns," in *Advances in Biometrics*. Springer, 2007, pp. 464–473.

[38] C. Shan, "Learning local binary patterns for gender classification on real-world face images," *Pattern Recognition Letters*, vol. 33, no. 4, pp. 431–437, 2012.

[39] V. Singh, V. Shokeen, and M. B. Singh, "Comparison of feature extraction algorithms for gender classification from face images," in *International Journal of Engineering Research and Technology*, no. 5 (May-2013). ESRSA Publications, 2013.

[40] G. Guo, C. R. Dyer, Y. Fu, and 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.

[41] J. Bekios-Calfa, J. M. Buenaposada, and L. Baumela, "Robust gender recognition by exploiting facial attributes dependencies," *Pattern Recognition Letters*, vol. 36, pp. 228–234, 2014.

[42] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[43] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *European conference on computational learning theory*. Springer, 1995, pp. 23–37.

[44] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 971–987, 2002.

[45] P. Dago-Casas, D. González-Jiménez, L. L. Yu, and J. L. Alba-Castro, "Single-and cross-database benchmarks for gender classification under unconstrained settings," in *Computer vision workshops (ICCV Workshops), 2011 IEEE international conference on*. IEEE, 2011, pp. 2152–2159.

[46] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[47] L. A. Alexandre, "Gender recognition: A multiscale decision fusion approach," *Pattern Recognition Letters*, vol. 31, no. 11, pp. 1422–1427, 2010.

[48] J. E. Tapia and C. A. Perez, "Gender classification based on fusion of different spatial scale features selected by mutual information from histogram of lbp, intensity, and shape," *Information Forensics and Security, IEEE Transactions on*, vol. 8, no. 3, pp. 488–499, 2013.

[49] N. Wang, X. Gao, D. Tao, and X. Li, "Facial feature point detection: A comprehensive survey," *arXiv preprint arXiv:1410.1037*, 2014.

[50] R. Brunelli and T. Poggio, "Face recognition: Features versus templates," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 10, pp. 1042–1052, 1993.

[51] Y. S. El-Din, M. N. Moustafa, and H. Mahdi, "Landmarks-sift face representation for gender classification," in *International*

Conference on Image Analysis and Processing. Springer, 2013, pp. 329–338.

[52] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[53] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 34–42.

[54] J. van de Wolfshaar, M. F. Karaaba, and M. A. Wiering, "Deep convolutional neural networks and support vector machines for gender recognition," in *Computational Intelligence, 2015 IEEE Symposium Series on*. IEEE, 2015, pp. 188–195.

[55] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *arXiv preprint arXiv:1603.01249*, 2016.

[56] S. Jia and N. Cristianini, "Learning to classify gender from four million images," *Pattern Recognition Letters*, vol. 58, pp. 35–41, 2015.

[57] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[58] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *European Conference on Computer Vision*. Springer, 2016, pp. 597–613.

[59] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2011.

[60] B. Gong, K. Grauman, and F. Sha, "Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation," in *International Conference on Machine Learning*, 2013, pp. 222–230.

[61] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang, "Domain adaptation under target and conditional shift," in *International Conference on Machine Learning*, 2013, pp. 819–827.

[62] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2066–2073.

[63] M. Baktashmotlagh, M. T. Harandi, B. C. Lovell, and M. Salzmann, "Domain adaptation on the statistical manifold," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2481–2488.

[64] ——, "Unsupervised domain adaptation by domain invariant projection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 769–776.

[65] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," *Computer Vision–ECCV 2010*, pp. 213–226, 2010.

[66] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: A survey of recent advances," *IEEE signal processing magazine*, vol. 32, no. 3, pp. 53–69, 2015.

[67] S. Banerjee, S. Samanta, and S. Das, "Face recognition in surveillance conditions with bag-of-words, using unsupervised domain adaptation," in *Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing*. ACM, 2014, p. 50.

[68] X. Tan, S. Chen, Z.-H. Zhou, and F. Zhang, "Face recognition from a single image per person: A survey," *Pattern recognition*, vol. 39, no. 9, pp. 1725–1745, 2006.

[69] F. Hafiz, A. A. Shafie, and Y. M. Mustafah, "Face recognition from single sample per person by learning of generic discriminant vectors," *Procedia Engineering*, vol. 41, pp. 465–472, 2012.

[70] B. Wang, W. Li, Z. Li, and Q. Liao, "Adaptive linear regression for single-sample face recognition," *Neurocomputing*, vol. 115, pp. 186–191, 2013.

[71] W. Deng, J. Hu, X. Zhou, and J. Guo, "Equidistant prototypes embedding for single sample based face recognition with generic learning and incremental learning," *Pattern Recognition*, vol. 47, no. 12, pp. 3738–3749, 2014.

[72] P. Zhu, M. Yang, L. Zhang, and I.-Y. Lee, "Local generic representation for face recognition with single sample per person," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 34–50.

[73] M. Yang, L. Van Gool, and L. Zhang, "Sparse variation dictionary learning for face recognition with a single training sample per person," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 689–696.

[74] M. Kafai, L. An, and B. Bhanu, "Reference face graph for face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2132–2143, 2014.

[75] X. Hu, W.-x. Yu, and J. Yao, "Multi-oriented 2dpca for face recognition with one training face image per person," *Journal of Computational Information Systems*, vol. 6, no. 5, pp. 1563–1570, 2010.

[76] H. Yin, P. Fu, and S. Meng, "Sampled flda for face recognition with single training image per person," *Neurocomputing*, vol. 69, no. 16, pp. 2443–2445, 2006.

[77] Y. Wang, M. Wang, Y. Chen, and Q. Zhu, "A novel virtual samples-based sparse representation method for face recognition," *Optik-International Journal for Light and Electron Optics*, vol. 125, no. 15, pp. 3908–3912, 2014.

[78] X. Chen and J. Zhang, "Illumination robust single sample face recognition using multi-directional orthogonal gradient phase faces," *Neurocomputing*, vol. 74, no. 14, pp. 2291–2298, 2011.

[79] R.-X. Ding, D. K. Du, Z.-H. Huang, Z.-M. Li, and K. Shang, "Variational feature representation-based classification for face recognition with single sample per person," *Journal of Visual Communication and Image Representation*, vol. 30, pp. 35–45, 2015.

[80] I. Kemelmacher-Shlizerman and R. Basri, "3d face reconstruction from a single image using a single reference face shape," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 2, pp. 394–405, 2011.

[81] D. Jiang, Y. Hu, S. Yan, L. Zhang, H. Zhang, and W. Gao, "Efficient 3d reconstruction for face recognition," *Pattern Recognition*, vol. 38, no. 6, pp. 787–798, 2005.

[82] C. Hu, M. Ye, S. Ji, W. Zeng, and X. Lu, "A new face recognition method based on image decomposition for single sample per person problem," *Neurocomputing*, vol. 160, pp. 287–299, 2015.

[83] A. T.-Y. Chen, M. Biglari-Abhari, K. I.-K. Wang, A. Bouzerdoum, and F. H. C. Tivive, *Hardware/Software Co-design for a Gender Recognition Embedded System.* Springer International Publishing, 2016, pp. 541–552.

[84] A. Ratnakar and G. More, "Real time gender recognition on fpga," *Perception*, vol. 6, pp. 19–22, 2015.

[85] B. He, D. Xu, R. Nian, M. van Heeswijk, Q. Yu, Y. Miche, and A. Lendasse, "Fast face recognition via sparse coding and extreme learning machine," *Cognitive Computation*, vol. 6, no. 2, pp. 264–277, 2014.

[86] Z. Zhang, W. Li, and H. Jia, "A fast face recognition algorithm based on mapreduce," in *Computational Intelligence and Design (ISCID), 2014 Seventh International Symposium on*, vol. 2. IEEE, 2014, pp. 395–399.

[87] G. Mahale, H. Mahale, A. Goel, S. Nandy, S. Bhattacharya, and R. Narayan, "Hardware solution for real-time face recognition," in *VLSI Design (VLSID), 2015 28th International Conference on.* IEEE, 2015, pp. 81–86.

[88] K. Selvakumar, J. Jerome, N. Shankar, and T. Sarathkumar, "Robust embedded vision system for face detection and identification in smart surveillance," *International Journal of Signal and Imaging Systems Engineering*, vol. 8, no. 6, pp. 356–366, 2015.

[89] A. Kushsairy, M. K. Kamaruddin, H. Nasir, S. I. Safie, Z. A. K. Bakti, M. R. Isa, and S. Khan, "Embedded vision: Enhancing embedded platform for face detection system," in *Instrumentation and Measurement Technology Conference Proceedings (I2MTC), 2016 IEEE International.* IEEE, 2016, pp. 1–5.

[90] BDTI, "Embedded vision alliance," Available: https://www.embedded-vision.com/, 2017.

[91] W. J. MacLean, "An evaluation of the suitability of fpgas for embedded vision systems," in *Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on.* IEEE, 2005, pp. 131–131.

[92] K. Pauwels, M. Tomasi, J. D. Alonso, E. Ros, and M. M. Van Hulle, "A comparison of fpga and gpu for real-time phase-based optical flow, stereo, and local image features," *IEEE Transactions on Computers*, vol. 61, no. 7, pp. 999–1012, 2012.

[93] S. Asano, T. Maruyama, and Y. Yamaguchi, "Performance comparison of fpga, gpu and cpu in image processing," in *Field programmable logic and applications, 2009. fpl 2009. international conference on.* IEEE, 2009, pp. 126–131.

[94] D. Bouris, A. Nikitakis, and I. Papaefstathiou, "Fast and efficient fpga-based feature detection employing the surf algorithm," in *Field-Programmable Custom Computing Machines (FCCM), 2010 18th IEEE Annual International Symposium on.* IEEE, 2010, pp. 3–10.

[95] M. Schaeferling and G. Kiefer, "Object recognition on a chip: a complete surf-based system on a single fpga," in *Reconfigurable Computing and FPGAs (ReConFig), 2011 International Conference on*. IEEE, 2011, pp. 49–54.

[96] M. Schaeferling, U. Hornung, and G. Kiefer, "Object recognition and pose estimation on embedded hardware: Surf-based system designs accelerated by fpga logic," *International Journal of Reconfigurable Computing*, vol. 2012, p. 6, 2012.

[97] M. Malik, F. Farahmand, P. Otto, N. Akhlaghi, T. Mohsenin, S. Sikdar, and H. Homayoun, "Architecture exploration for energy-efficient embedded vision applications: From general purpose processor to domain specific accelerator," in *VLSI (ISVLSI), 2016 IEEE Computer Society Annual Symposium on*. IEEE, 2016, pp. 559–564.

[98] F. Conti, D. Rossi, A. Pullini, I. Loi, and L. Benini, "Pulp: A ultra-low power parallel accelerator for energy-efficient and flexible embedded vision," *Journal of Signal Processing Systems*, vol. 84, no. 3, pp. 339–354, 2016.

[99] B. Zhang, C. Zhao, K. Mei, N. Zheng *et al.*, "Hierarchical and parallel pipelined heterogeneous soc for embedded vision processing," *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[100] S. Mattoccia, I. Marchio, and M. Casadio, "A compact 3d camera suited for mobile and embedded vision applications," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 195–196.

[101] V. Carletti, L. Del Pizzo, G. Percannella, and M. Vento, "Foreground detection optimization for socs embedded on smart cameras," in *Proceedings of the International Conference on Distributed Smart Cameras*, ser. ICDSC '14. New York, NY, USA: ACM, 2014, pp. 31:1–31:5. [Online]. Available: http://doi.acm.org/10.1145/2659021.2659060

[102] S. Ehsan, A. F. Clark, K. D. McDonald-Maier *et al.*, "Integral images: efficient algorithms for their computation and storage in

resource-constrained embedded vision systems," *Sensors*, vol. 15, no. 7, pp. 16 804–16 830, 2015.

[103] A. Suleiman, Y.-H. Chen, J. Emer, and V. Sze, "Towards closing the energy gap between hog and cnn features for embedded vision," *arXiv preprint arXiv:1703.05853*, 2017.

[104] C. Hu, F. Arvin, C. Xiong, and S. Yue, "A bio-inspired embedded vision system for autonomous micro-robots: the lgmd case," *IEEE Transactions on Cognitive and Developmental Systems*, 2017.

[105] G. Velez, A. Cortés, M. Nieto, I. Vélez, and O. Otaegui, "A reconfigurable embedded vision system for advanced driver assistance," *Journal of Real-Time Image Processing*, vol. 10, no. 4, pp. 725–739, 2015.

[106] P. Foggia, A. Greco, A. Saggese, and M. Vento, "A method for detecting long term left baggage based on heat map." in *VISAPP (2)*, 2015, pp. 385–391.

[107] V. Carletti, P. Foggia, A. Greco, A. Saggese, and M. Vento, "Automatic detection of long term parked cars," in *Advanced Video and Signal Based Surveillance (AVSS), 2015 12th IEEE International Conference on*. IEEE, 2015, pp. 1–6.

[108] L. DelPizzo, P. Foggia, A. Greco, G. Percannella, and M. Vento, "Counting people by {RGB} or depth overhead cameras," *Pattern Recognition Letters*, vol. 81, pp. 41 – 50, 2016.

[109] V. N. Vapnik, "An overview of statistical learning theory," *IEEE transactions on neural networks*, vol. 10, no. 5, pp. 988–999, 1999.

[110] L. Lam and C. Y. Suen, "Optimal combinations of pattern classifiers," *Pattern Recognition Letters*, vol. 16, no. 9, pp. 945–954, 1995.

[111] J. Kittler, "Combining classifiers: A theoretical framework," *Pattern analysis and Applications*, vol. 1, no. 1, pp. 18–27, 1998.

[112] P. Soda, G. Iannello, and M. Vento, "A multiple expert system for classifying fluorescent intensity in antinuclear autoantibodies analysis," *Pattern Analysis and Applications*, vol. 12, no. 3, p. 215, 2009.

[113] M. De Santo, M. Molinara, F. Tortorella, and M. Vento, "Automatic classification of clustered microcalcifications by a multiple expert system," *Pattern Recognition*, vol. 36, no. 7, pp. 1467–1477, 2003.

[114] L.-L. Huang and A. Shimizu, "A multi-expert approach for robust face detection," *Pattern Recognition*, vol. 39, no. 9, pp. 1695–1703, 2006.

[115] L. P. Cordella, M. De Santo, G. Percannella, C. Sansone, and M. Vento, "A multi-expert system for movie segmentation," *Lecture notes in computer science*, pp. 304–313, 2002.

[116] M. Uricár, V. Franc, and V. Hlavác, "Facial landmark tracking by tree-based deformable part model based detector," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 10–17.

[117] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*. Springer, 2006, pp. 404–417.

[118] G. Azzopardi and N. Petkov, "Trainable COSFIRE filters for keypoint detection and pattern recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 2, pp. 490–503, 2013.

[119] ——, "Ventral-stream-like shape representation: from pixel intensity values to trainable object-selective cosfire models," *Front Comput Neurosci*, vol. 8, p. 80, 2014. [Online]. Available: http://www.biomedsearch.com/nih/Ventral-stream-like-shape-representation/25126068.html

[120] B. Gecer, G. Azzopardi, and N. Petkov, "Color-blob-based COSFIRE filters for object recognition," *Image and Vision*

Computing, vol. 57, pp. 165 – 174, 2017. [Online]. Available: // www.sciencedirect.com/science/article/pii/S0262885616301895

[121] G. Azzopardi, N. Strisciuglio, M. Vento, and N. Petkov, "Trainable cosfire filters for vessel delineation with application to retinal images," *Medical Image Analysis*, vol. 19, no. 1, p. 46?57, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361841514001364

[122] N. Strisciuglio, G. Azzopardi, M. Vento, and N. Petkov, "Supervised vessel delineation in retinal fundus images with the automatic selection of B-COSFIRE filters," *Machine Vision and Applications*, p. 1?13, 2016. [Online]. Available: http://dx.doi.org/10.1007/s00138-016-0781-7

[123] G. Azzopardi and N. Petkov, "A CORF computational model of a simple cell that relies on lgn input outperforms the gabor function model," *Biological Cybernetics*, vol. 106, pp. 177–189, 2012, 10.1007/s00422-012-0486-6. [Online]. Available: http://dx.doi.org/10.1007/s00422-012-0486-6

[124] G. Azzopardi, A. Rodríguez-Sánchez, J. Piater, and N. Petkov, "A push-pull CORF model of a simple cell with antiphase inhibition improves snr and contour detection," *PLoS ONE*, vol. 9, no. 7, p. e98424, 07 2014. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pone.0098424

[125] A. Pasupathy and C. E. Connor, "Population coding of shape in area V4," *Nature Neuroscience*, vol. 5, no. 12, pp. 1332–1338, DEC 2002.

[126] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, ser. CVPR '06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 2169–2178. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2006.68

[127] S. Tulyakov, S. Jaeger, V. Govindaraju, and D. Doermann, "Review of classifier combination methods," *Machine Learning in Document Analysis and Recognition*, pp. 361–386, 2008.

[128] R. DiLascio, P. Foggia, G. Percannella, A. Saggese, and M. Vento, "A real time algorithm for people tracking using contextual reasoning," *Computer Vision and Image Understanding*, vol. 117, no. 8, pp. 892 – 908, 2013.

[129] RAPP, "Raster processing primitives," Available: https://savannah.nongnu.org/projects/rapp, 2017.

[130] E. Zhou, Z. Cao, and Q. Yin, "Naive-deep face recognition: Touching the limit of LFW benchmark or not?" *CoRR*, vol. abs/1501.04690, 2015. [Online]. Available: http://arxiv.org/abs/1501.04690

[131] Luxand, "Luxand api," https://www.luxand.com/download/Luxand_FaceSDK_Documentation.pdf, 2017.

[132] BetaFace, "Beta face api," https://www.betaface.com/wpa/wp-content/uploads/2014/01/Betaface-SDK.pdf, 2017.

[133] Kairos, "Kairos human analytic sdk," https://www.kairos.com/docs/sdk, 2017.

[134] MicrosoftFace, "Microsoft face api," https://dev.projectoxford.ai/docs/services/563879b61984550e40cbbe8d/operations/563879b61984550f30395236, 2017.