UNIVERSITÀ
DEGLI STUDI
FIRENZE
DINFO
DIPARTIMENTO DI
INGEGNERIA
DELL'INFORMAZIONE

DOTTORATO DI RICERCA IN
INFORMATICA, SISTEMI E TELECOMUNICAZIONI
MEDIA INTEGRATION AND COMMUNICATION CENTER (MICC)
ING-INF/05

# MULTI-TARGET TRACKING AND FACIAL ATTRIBUTE ESTIMATION IN SMART ENVIRONMENTS

*Candidate*
Dario Di Fina

*Supervisors*
Prof. Alberto Del Bimbo

Prof. Andrew D. Bagdanov

*PhD Coordinator*
Prof. Luigi Chisci

CICLO XXVIII, 2012-2015

Università degli Studi di Firenze, Dipartimento di Ingegneria dell'Informazione,
Media Integration and Communication Center (MICC).

Thesis submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computer Science, Systems and Telecommunications.

*Ai miei Genitori, a mio Fratello:*
*Questa tesi è mia quanto Loro.*
*Ai miei Nonni che sono sempre con me.*

*"Ogni focu forti cìnniri addiventa.*
*Ogni principio è forte, e ogni desiderio vene a fine."*

# Acknowledgments

*A good head and a good heart are always a formidable combination.*

- Nelson Mandela.

Now it's time! After these last three years I finally completed my PhD thesis. This period taught me so many things that I could write another thesis, but there is no time! Now I want to thank all those people who made possible this final work.

First of all, I would like to express my sincere gratitude to my supervisor. Professor Alberto Del Bimbo for the opportunity that he gave me, and for his precious advices during this three years. I also would like to thank Professor Andrew D. Bagdanov for his valuable help and his continuous support he guided me, collaborated with me on the main parts of my work and taught me how to approach and develop my research activity.

My thanks go to Dr. Giuseppe Lisanti and Dr. Svebor Karaman, which substantially contributed to my research. I thank all my colleagues of the Media Integration and Communication Center (MICC) who supported me during my research: my mates Rudy Becarelli and Tiberio Uricchio with whom we begun this journey, Prof. Marco Bertini, Dr. Lorenzo Seidenari, Dr. Federico Pernici, Claudio Baecchi, Enrico Bondi, Federico Becattini, Maxime Devanne, Leonardo Galteri, Francesco Gelli, Francesco Turchini, Dr. Lamberto Ballan, Andrea Ferracani, Daniele Pezzatini, Dr. Lea Landucci, Federico Bartoli, Simone Ercoli, Joël Dumoulin, Professor Pala, Professor Berretti, Matteo Casini, Dr. Roberto Caldelli, Dr Irene Amerini, and the other fellows that I met at MICC.

Heartfelt thanks to all my family, my aunt Santina, my uncle Salvatore, my cousins Gandolfo, Massimo, Gianluca, Fabio, my aunt Concetta and my cousins Salvatore and Lucia, my aunts Angela and Gandolfa and my uncle Gianni, my cousins Michelangelo, Pierluigi and Laura, all their families and all my relatives who are not with us anymore, and won't be able to see my graduation, but I know that they will be watching me from above, you will always be with me!

I also thank all my friends: first of all Elisa who helped me enormously in these days, Rino, Randi, Stefano, Mattia, Silvia, Vittoria and their family, Erica, Sara G., Giacomo, Francesco R., Sara B., Francesco D., Caterina, Luca, Alessandra, Silvia, Duba, David, Manuel, Diego, Chiara Selene, Ruggero, Kevin, Matteo and all my mates of the dojo, Alessandro, Valentina, Niccolò, Lisa, Matteo, Francesca, Federico, Sergio and Fabio. Last but not least I would like to thank all those people I met in these years and treated me for who I really am.

A big thank to whom will read this thesis or even just leaf through it.

Again, my special heartfelt thanks goes to Stefania, Francesco and Riccardo! Without them I would have never completed this journey.

Thank you all!!!!

# Abstract

This dissertation presents a study on three different computer vision topics that have applications to smart environments. We first propose a solution to improve multi-target data association based on $\ell_1$-regularized sparse basis expansions. The method aims to improve the data association process by addressing problems like occlusion and change of appearance. Experimental results show that, for the pure data association problem, our proposed approach achieves state-of-the-art results on standard benchmark datasets.

Next, we extend our new data association approach with a novel technique based on a weighted version of sparse reconstruction that enforces long-term consistency in multi-target tracking. We introduce a two-phase approach that first performs local data association, and then periodically uses accumulated usage statistics in order to merge tracklets and enforce long-term, global consistency in tracks. The result is a complete, end-to-end tracking system that is able to reduce tracklet fragmentation and ID switches, and to improve the overall quality of tracking.

Finally, we propose a method to jointly estimate face characteristics such as Gender, Age, Ethnicity and head pose. We develop a random forest based method based around a new splitting criterion for multi-objective estimation. Our system achieves results comparable to the state-of-the-art, and has the additional advantage of simultaneously estimating multiple facial characteristics using a single pool of image features rather than characteristic-specific ones.

# Contents

**Abstract** **vii**

**Contents** **ix**

**1 Introduction** **1**

1.1 Objectives . . . 1

1.2 Our Contributions . . . 8

1.3 Organization of the Thesis . . . 9

**2 Literature Review** **11**

2.1 Multiple Target Tracking . . . 12

2.1.1 Data Association . . . 13

2.1.2 Tracking Algorithm . . . 14

2.1.3 Tracking-by-detection . . . 15

2.2 Facial Characteristic Estimation . . . 16

2.2.1 Gender Classification . . . 16

2.2.2 Age Estimation . . . 17

2.2.3 Soft Biometric Estimation . . . 17

2.2.4 Head Pose Estimation . . . 18

**3 Data Association using $\ell_1$-regularization** **21**

3.1 Sparse Reconstruction and Sparse Coding . . . 23

3.2 Data Association by Sparse Reconstruction . . . 25

3.2.1 The Data Association Problem . . . 26

3.2.2 Sparse Discriminative Basis Expansion for Data Association 26

3.2.3 Data Association Algorithm . . . 30

3.3 Experiments . . . 32

3.3.1 Data Association Performance . . . . . . . . . . . . . . . . 32
3.3.2 Comparison with the State-of-the-art . . . . . . . . . . . . . 33
3.4 Discussion and Conclusions . . . . . . . . . . . . . . . . . . . . . 34

**4 Multi-Target Tracking using the Weighted Sparse Reconstruction 37**
4.1 Multi-Target Tracking by Sparse Reconstruction . . . . . . . . . . . 39
4.1.1 The Multi-Target Tracking Problem . . . . . . . . . . . . . 39
4.1.2 Local Data Association . . . . . . . . . . . . . . . . . . . . 40
4.1.3 Global Association for Long-term Consistency . . . . . . . 45
4.2 Experimental Results . . . . . . . . . . . . . . . . . . . . . . . . . 50
4.2.1 Datasets and Metric . . . . . . . . . . . . . . . . . . . . . . 51
4.2.2 Performance on PETS . . . . . . . . . . . . . . . . . . . . 52
4.2.3 Performance on ETH . . . . . . . . . . . . . . . . . . . . . 53
4.3 Discussion and Conclusions . . . . . . . . . . . . . . . . . . . . . 54

**5 Multi-Objective Random Forests for Multiple Facial Characteristic Estimation 55**
5.1 Random Forests . . . . . . . . . . . . . . . . . . . . . . . . . . . 57
5.1.1 Random Forests for Face Characteristic Estimation . . . . . 60
5.2 Multi-objective Random Forests (MORF) . . . . . . . . . . . . . . 61
5.2.1 Feature Representation for Faces . . . . . . . . . . . . . . . 61
5.2.2 Random Forests for Supervised Estimation . . . . . . . . . 61
5.3 Experimental Results . . . . . . . . . . . . . . . . . . . . . . . . . 65
5.3.1 Datasets and Experimental Protocols . . . . . . . . . . . . . 65
5.3.2 Analysis of Multi-objective Learning . . . . . . . . . . . . 69
5.3.3 Comparison with the State-of-the-art . . . . . . . . . . . . . 70
5.4 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 73

**6 Conclusions 75**
6.1 Concluding Remarks . . . . . . . . . . . . . . . . . . . . . . . . . 75
6.2 Summary of Contributions . . . . . . . . . . . . . . . . . . . . . . 76
6.3 Expected Impact of the Proposed Research . . . . . . . . . . . . . 77

**A Publications 79**

**Bibliography 81**

# Chapter 1

# Introduction

*We are not powerless. We have tremendous potential for good or ill. How we choose to use that power is up to us; but first we must choose to use it. We're told every day, "You can't change the world." But the world is changing every day. Only question is...who's doing it? You or somebody else?*

- J. Michael Straczynski.

## 1.1 Objectives

In the last decade, there has been a steady growth in video surveillance applications due to the necessity to improve security systems in sensitive places like airports, train stations or crowded places. This increase has been made possible by continuous technology improvement in terms of video cameras and computing systems, allowing the creation of portable or wearable devices. This has led to a greater interest in this research topic, defining and proposing new issues, such as the analysis of human behavior [19, 79], interactions in groups of people [43] (in Figure 1.2(a) and in Figure 1.2(b) two examples of interactions are shown), or flow analysis of large crowds [33] for the identification of abnormal events (for example in Figure 1.1(a) [6] and in Figure 1.1(b) [1]). All these factors have ultimately led the scientific community to formalize the concept of *Smart Cities*, aimed at the realization of urban environments able to work automatically to improve quality of life, to guarantee the safety of its citizens, and to guarantee special services also to visitors and tourists.

(a)

(b)

Figure 1.1: Examples of two different crowded scenes. (a) a frame acquired from a surveillance camera in a city. (b) a frame acquired on a university campus.

In general the term *smart environments* includes several concepts like smart cities, smart houses and smart museums. It refers to physical spaces equipped with sensors feeding into adaptive algorithms that enable the environment to become sensitive and responsive to the presence and needs of its occupants. Smart environment applications range from navigating through an unfamiliar space, to providing context-sensitive reminders for activities. In order to support its occupants, the smart environment must be able to detect the environment, outdoor or indoor, detect people and the actions they perform, all the while acquiring relevant information from each of them.

Information extracted can include physical gestures, relationships between people and objects, features of the physical environment such as spatial layout and temperature, or person identity and localization. For example, in [71] the authors propose a method for face recognition in smart environments, and in [39] the authors define a tracking system for realistic home environments. In both cases the authors propose an approach to analyze specific issues in real application scenarios.

In the context of this dissertation we consider computer vision techniques and technologies that can be used to support smart environment applications. In particular, we approach three specific topics:

- **the multiple data association problem in visual tracking** for which we have devised a system to robustly associate multiple target in an ideal tracking scenario (i.e. noise-free);

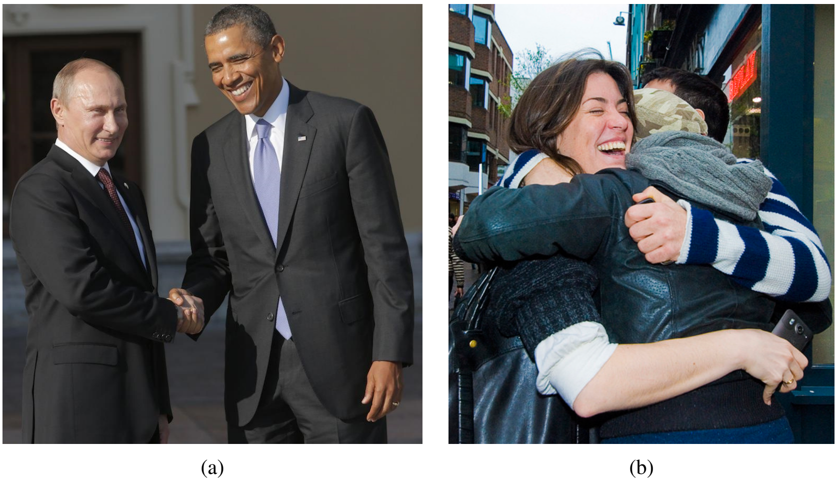
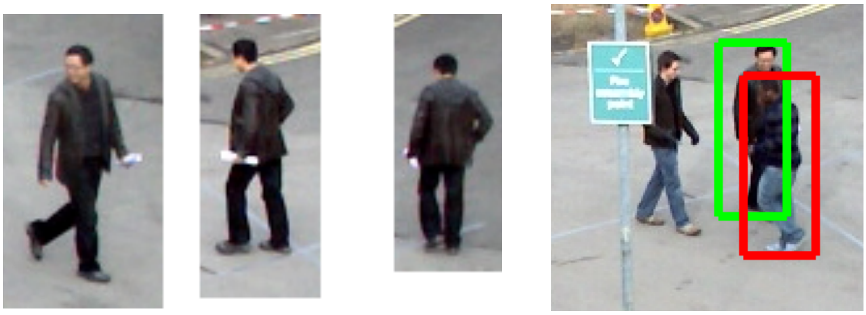


Figure 1.2: Examples of interaction between people. (a) an example of handshake between two people. (b) an informal situation of three people hugging.

- **the end-to-end multi-target tracking problem** for which we propose a global data association method to ensure tracking consistency over several targets within a video scene; and
- **the facial characteristic estimation problem** for which we introduce a new system capable of jointly estimating biometric facial characteristics of a subject.

All three topics are pertinent to video surveillance applications, in fact thanks to a customized and content based profiling, specific and selected multimedia could be proposed according to the estimated characteristics. In terms of security applications, one could think of a security company that automatically monitors a wide area that can verify instantly the identity of a given person, analysing each movement and the relevant face characteristics.

In this dissertation we examine the above mentioned topics applied to video surveillance systems. We will first introduce the three topics considering the issues that need particular attention for each of them. The multi-target data association problem is one of the main challenging one in visual tracking and consists of finding the correct assignment between the set of tracked targets and the set of new observations extracted from the current frame of a sequence. Some of the main complex issues related to this problem are occlusion and identity switches. If one considers



(a) (b)

Figure 1.3: Difficult situations for trackers. (a) the same target with three different appearances in a single video sequence. (b) an example of occlusion between two targets. Both of these phenomena are common in surveillance video and render tracking difficult.

a complete multi-target tracking process there may be problems like tracklet fragmentation or missed detections, which hinder accurate tracking of multiple subjects in the video scene. Facial characteristic estimation analyzes images of people in order to estimate specific properties such as Age, Gender or Ethnicity. The main complicating factor in this problem is the fact that the selected technique should be robust to head pose variations of the subjects. As a consequence most systems define a customized approach and task-specific descriptors for each particular face characteristic.

In the following sections we introduce in detail each of the topics considered in this thesis and we sketch our proposed solutions for each of them as a preview of the main technical contributions.

### Multi-Target Tracking

The visual tracking process is the localization of one or more subjects that are labelled and collected through time in a video scene (in Figure 1.4 a tracking process of four targets is illustrated). Each step in this process has to be handled and interconnected with the others. Under ideal conditions, during the tracking process the system should estimate for each time instant all target observations even when they are only partly visible. However, in real conditions many problems such as ID switches, missed detections, tracklet fragmentation, illumination changes and

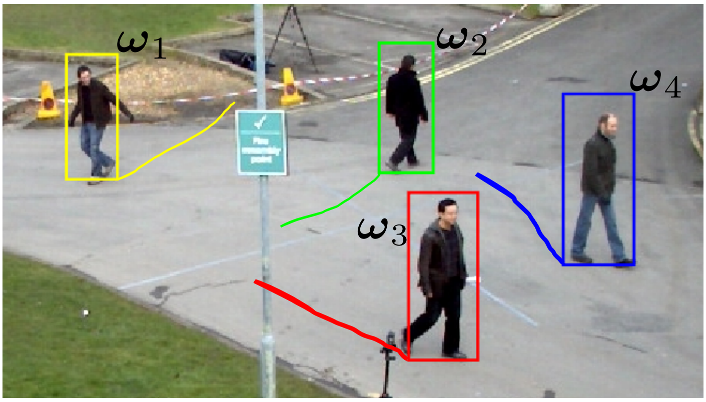


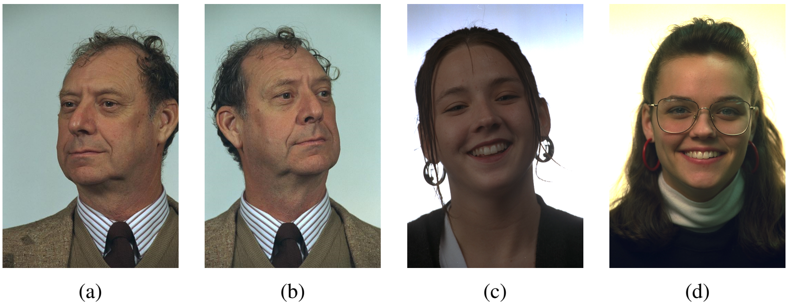
Figure 1.4: Tracking four targets. The colored lines represent the path followed by the targets in the video sequence.

change of appearance must be surmounted simultaneously (e.g. Figure 1.3(a)).

Occlusions are cases which occur when during the tracking process a target is partially or completely occluded by scene objects or by other subjects (e.g. Figure 1.3(b)). Occlusions can be controlled in different ways, for example considering a prediction model that use spatio-temporal information to anticipate the next position of each target. Another way to handle this problem is to use a part-based model for person detection, as in [66, 84], in order to maintain a model robust to occlusion (only some parts will be occluded at any time).

When multiple targets are present, occlusion may cause identity switches. Identity switches happen when, in the tracking process, the observation of a particular subject is associated incorrectly with another one, producing wrong associations and decreasing the accuracy of the approach. An identity switch can occur in different ways, but mainly because of occlusions during detection, or because of the presence of another subject with similar appearance. In cases of occlusion, if a robust model is used, the ID switch is restricted to the frames in which the target is occluded. However, when targets with similar appearance are present it is necessary to consider other factors to resolve data association ambiguities, using spatial information, or defining a score [3, 94] to evaluate the similarity between observations and targets.

Tracklet fragmentation is yet another challenge in multi-target tracking. It occurs in cases in which ID switches are not limited to a few frames, or in ones in which the same targets are labelled with different identities. One (expensive) solu-



(a) (b) (c) (d)

Figure 1.5: Examples of face image acquisition. In (a) and (b) the same person with two different face orientation is shown, in (c) an image acquired with a low brightness, and in (d) a girl wearing glasses. Image variability like this is one of the main challenges that must be addressed for robust facial characteristic estimation.

tion is to evaluate tracklet pairs as candidates for merging. Alternatively, a global process that uses the target information to verify possible relationships between tracklets can be applied.

Addressing each of these issues in multi-target tracking requires different considerations and techniques, and each of them is related to a different sub-problem in the visual tracking process. In Chapter 3 of this dissertation we propose an innovative multi-target data association algorithm to handle occlusions and ID switch issues. It defines and exploits robust and discriminative models robust to appearances changes. In Chapter 4 we also define and end-to-end multi-target tracking system with a global association process that improves the long-term consistency, and decreases tracklet fragmentation and ID switches.

### Face Characteristic Estimation

Facial characteristics are used daily by our brain to recognize our friends and family, to categorize each of them almost instantly. We learn to classify each characteristic from childhood, accumulating information and improving our classification. Analysis of facial characteristics has gained more and more interest in recent years, not only for security reasons, but also in cultural studies. However, despite the constant evolution in techniques and instruments many unresolved issues remain.

One of the main problems in this research topic is that depending on the face orientation (in Figure 1.5(a) and 1.5(b) two different orientations of the same face

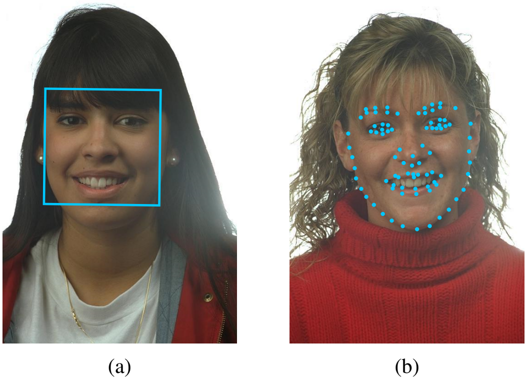


Figure 1.6: An example of face detection (a) and facial landmark estimation (b). Face detection is an essential precursor to face characteristic estimation, and many techniques exploit facial landmarks in order to extract targeted features for characteristic estimation.

image are reported), the estimation of a characteristic rather than another can be difficult, not allowing a properly acquisition of all relevant characteristics. In this case, for example, a training phase that uses not only face images with frontal pose, but also images with different pose, both pan and tilt angles can be considered. To facilitate this process, some works [89, 103] train their own method with different subsets, where each of them considers a specific range of angles. In some cases facial landmarks [105] are used: they allow to improve the characteristic estimation or simply enhance face detection. The facial landmarks are points that locate eyes, mouth or jaw (e.g. Figure 1.6(b)), which improve detection in those situations in which part of the face is occluded by an object or the subject is simply not frontal. Moreover, landmarks allow to extract from faces more discriminative information using a classification approach.

Another issue that is important to consider in any face analysis process is illumination change (e.g. Figure 1.5(c)). In a many cases during the acquisition process, face images can be acquired with different illumination, or with different objects like glasses (e.g. Figure 1.5(d)) or hats. In addiction somatic traits relevant to different ethnic groups are also complicating factors. For this reason it is necessary to define more discriminative descriptors, and in some cases to maintain separate techniques to estimate the most important characteristics, such as Gender or Age.

More often in different approaches descriptors combining multiple types of feature are used.

To improve facial characteristic estimation performance, face images are sometimes acquired as 3D models [63, 64]. This enables analysis in terms of traits that cannot be measured two-dimensional images. This typically requires that one or more images be acquired from which a 3D face model can be reconstructed. This can be quite complicated in non-cooperative use case scenarios where target motion is unrestricted. Moreover, for smart environments 3D acquisition may not be feasible due to cost and environmental constraints. Invasive imaging devices are not fully compatible with smart environments.

It appears evident, however, that facial characteristic estimation in most systems is designed to solve specific problems. This led us to study ways to estimate multiple characteristics because in real applications it can be much more efficient to estimate them jointly, using the same descriptor rather than separate them for each purpose. Such a system would also speed up the per-frame estimation process, which is essential in real-time systems. For this reason we have designed a way to jointly estimate multiple facial characteristic with the idea to define a method capable of estimating face pose, Age, Ethnicity and Gender.

## 1.2 Our Contributions

Despite the attention received in last years, the research topics considered in this dissertation remain extremely challenging with many unsolved issues. For multiple target tracking, two different problems have been studied: the multi-target data association and the long-term consistency problem. We approach the multi-target data association problem using a sparse reconstruction method robust to occlusion and change of appearance. We propose a new non-probabilistic approach that uses information accumulated over all local data associations to define a discriminative model for each target.

We then extend our approach to data association to address the long-term consistency issue. We define a new approach based on a weighted version of sparse reconstruction that can merge tracklets corresponding to the same target. This is a novel approach that does not consider statistical hypothesis to predict the location of each subject, but uses the collected data to handle at the same time several targets and decrease tracklet fragmentation.

With respect to facial characteristic estimation, existing approaches normally focus on a single characteristic estimation, or consider multiple characteristics independently. Random forests have been shown to be a powerful tool for head pose estimation [27,44], and we believe that they can be naturally extended to simultaneously estimation of multiple heterogeneous properties (including soft biometrics) of face imagery. We propose a random forest approach that jointly estimates multiple facial characteristics: Age, Gender, Ethnicity and head pose.

## 1.3 Organization of the Thesis

The overall structure of the thesis is herewith presented.

### Chapter 2: Literature Review

In Chapter 2 we propose an extensive review of the state-of-the-art methods for target tracking and facial characteristic estimation problem. For target tracking, the relevant data association approaches based on the Bayesian and other machine learning method are presented. The multi-target tracking method and in particular the tracking-by-detection technique is also reviewed. For face characteristic estimation , the recent works on Gender, Age and soft biometric estimation are discussed, illustrating how most methods are defined to handle separately the estimation of each of them.

### Chapter 3: Data Association using $\ell_1$-regularization

In Chapter 3 we present a multi-target data association method robust to occlusions and appearance changes. The proposed method is based on sparse reconstruction approach considering the $\ell_1$-regularization technique that will be introduced in Chapter 2. In order to properly study only the data association problem, the proposed approach was verified considering certain ideal conditions. In this context, ideal observations were used such that in each frame a pedestrian detector returns the correct detections that identify all targets present, even in the case of partial or total occlusions. In this way, we have been able to study pure data association problem without considering issues like miss detections, or false detections. A discriminative model able to robustly represent each identified target is also defined. This approach yields good results and we believe it is a good starting point for any tracking-by-detection technique.

### Chapter 4: Multi-Target Tracking using the Weighted Sparse Reconstruction

In Chapter 4 we extend the discriminative model and data association approach defined in Chapter 3 to an end-to-end multi-target tracking-by-detection system. It uses two versions of sparse reconstruction, and unlike the approach proposed in Chapter 3, assumes no ideal conditions. A more discriminative descriptor has been therefore selected to define a more robust target model. Two data associations phases are used. The first performs a frame-by-frame association between new observations and identified subjects, while the second one is used to increase the long-term consistency and decrease tracklet fragmentation and ID switches. We have tested the proposed technique on two benchmark datasets obtaining state-of-the-art results in both cases.

### Chapter 5: Multi-Objective Random Forests for Multiple Facial Characteristic Estimation

In Chapter 5 we use the random forests for the joint estimation of Gender, Age, Ethnicity and head pose. At the basis of the definition of this approach, we define a method that can simultaneously estimate multiple facial characteristics in order to simplify an estimation which in many cases is more complex than necessary. The proposed method is evaluated on publicly available imagery obtaining state-of-the-art results.

### Chapter 6: Conclusions

Chapter 6 summarizes the contribution of the thesis and proposes some further research activities on these topics. In the Appendix all published papers associated with this Ph.D. thesis are reported.

# Chapter 2

# Literature Review

> *Education never ends, Watson. It is a series of lessons, with the greatest for the last.*
>
> -Arthur Conan Doyle, *His Last Bow.*

In this chapter we give a brief review of available works related to multi-target tracking and face characteristic estimation methods. Both topics play a critical role in several applications, particularly in real world operation monitoring, video surveillance [2,4] and human–computer interaction [49]. The most significant state-of-the-art techniques are illustrated and the methods on which the proposed approaches have been based are finally introduced.

The multi-target tracking and the facial characteristic estimation problems are receiving a growing attention, while difficult problems remain and new research are carried out by the computer vision community. As far as visual tracking topics are concerned, a vast literature is available however, this is one of the main challenging problem. One of the main reasons for this is the uncontrolled nature of videos in which targets must be tracked and the necessity of coping with hard visual phenomena, such as uncontrolled and variable illumination that affects target appearance, target occlusion, unpredictable target motion, and the need to track a large number of target and very crowded scenes. All these factors have led to visual tracking remaining a very hot topic in the computer vision community.

Independently from these factors related to the application scenario that a tracker must surmount, the main difficulty is represented by *two* problems. Firstly, a multi-target tracker must be able to identify and track *individual* targets with robustness and precision (i.e. in other words the *visual target tracking problem* must

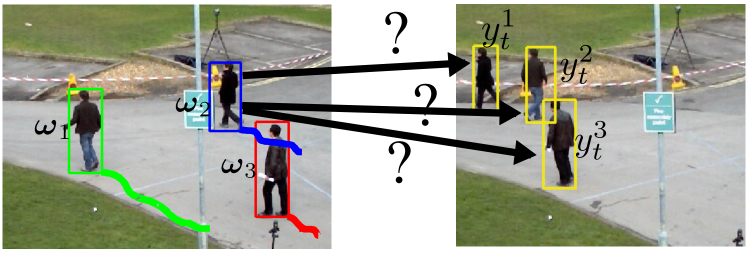


Figure 2.1: The data association problem. At each time instant observations $y_t^i$ must be associated with one of the running trackers $\omega_k$.

be first solved). This typically involves target detection, appearance modeling, target motion estimation, and location prediction in each frame. Secondly, the *data association problem* has to be faced. This mainly pertains the association of detected targets with one of the multiple tracked targets at each frame. Each one of these two problems is difficult in its own, and solving both of them simultaneously is therefore a great challenge.

Facial characteristics like Gender, Age, Ethnicity and face pose are important aspects to evaluate in many computer vision applications. Estimating Gender and Age can be a way to adapt advertising displayed on nearby screens, and pose estimation can allow users to interact with devices by simply looking at them. Although head pose is not a biometric characteristic, it is related to the gaze of a person and therefore can be an important aspect of behaviour and social interaction understanding. Despite the attention these applications have received in recent years, estimation of multiple face characteristics, and especially multiple soft biometrics such as Age, Gender and Ethnicity, remains an unsolved problem and an active area of research in the computer vision community.

In the following, we review some of the most relevant works on data association, tracking-by-detection and face characteristic estimation.

## 2.1 Multiple Target Tracking

As reported in the previous section, the multi-target tracking approach consists of different sub-problems each of which has its own challenges and peculiarities. Two of the most significant classes of tracking approach are those ones based on Bayesian filter [5, 10, 16, 29, 98] and those ones based on tracking-by-detection [9, 66, 70, 84]. The Bayesian filters approaches estimate the future targets state considering a recursive statistic model. The tracking-by-detection approach is more commonly used today, and in these cases the tracking algorithm uses the current frame and current targets information, defining a data association approach to perform the right association.

Here below the most significant approaches on data association and tracking algorithm can be found together with some details for the tracking-by-detection methods.

### 2.1.1 Data Association

In this section we review some of the most relevant works on data association methods. It is first of all observed that they can be addressed in a variety of ways (in Figure 2.1 an example of data association process is reported). A simple and widely-applied approach to multi-target data association is the Nearest Neighbour Standard Filter (NNSF) [5, 16, 28, 61]. It uses the Mahalanobis distance to compute the association error of a running tracker with observations at time $t$. The association process chooses the smallest Mahalanobis distance between all tracker/measurement pairs, iterating this process in a greedy way until all tracklets have been assigned to an observation. However the NNSF is susceptible to integration of incorrect measurements and can produce overconfident estimates over time.

In contrast to taking greedy local decisions about tracklet/observation associations, the authors of [101] propose a network flow based optimization method for data association. The approach considers a min-cost flow algorithm defining a global optimal trajectory that reduces tracklet fragmentation. The authors in [51] propose instead an approach based on a combination between Rao-Blackwellized Monte Carlo data association algorithm and the particle Markov chain Monte Carlo method to define a data association approach. On the other hand, in [47] the authors modify the Hungarian algorithm to manage occlusion and missed detection. Furthermore, they introduce a kind of node graph to handle different types of association errors.

With a Bayesian approach based on Switching Linear Dynamical System (SLDS) [67] the authors in [82] define an alternative parametrization of the data association problem, introducing a latent data association approach. They propose a data association technique which number of targets is implicitly inferred together with the association process and consider a single inference problem to solve data association and model selection using the SLDS. This inference-based approach incorporates inference over inliers/outliers and track termination times into the system.

Data association can often be decomposed into two phases: a *local association* phase and a *global association* phase. Generally, the local data association approach defines an association score between tracklets and observations, while global association verifies associations using a measure of tracklet similarity (and thus considering longer sequences). In [83] the authors define a local association approach using a tensor to model the interaction between tracklet and detection. They use a non-convex optimization formulation to find good solutions. This is obtained by computing context information used to verify interactions between tracklets and reduce fragmentation. On the other hand, the authors in [92] use a non-linear motion map and a sliding window to compute the local association between tracklets and new detection, and use an affinity measure to compute the global association process.

### 2.1.2 Tracking Algorithm

Target tracking is usually performed using a Kalman filter or other Bayesian filter that maintains a statistical model of target motion at each time step. Bayesian approaches exploit prior association probabilities and therefore allow to explicitly model measurement uncertainty. In this context, one of the most widely used technique is the Joint Probabilistic Data Association Filter (JPDAF) [16,17,28,29,98]. The JPDAF produces a set of hypotheses that associate tracklets with new observations, applies a gate to reduce the set of admissible hypotheses, and then computes a MAP estimate of the Bayes-optimal solution to the data association problem at the current time step. This method does not scale well with the number of targets and observations, since many hypotheses can be generated and this can result in a huge increase of the computational complexity of the MAP solution.

Another technique widely used in literature is Markov Chain Monte Carlo Data Association (MCMCDA) [23, 69]. This approach uses a Monte Carlo sampling process defined over “moves” that change associations over short intervals of time. By randomly sampling changes in the association according to a statistical model, the MCMCDA approach is able to efficiently search in a very large space and find a good approximation of the optimal data association.

In addition to the more commonly applied methods, different techniques that use learning algorithm to define new tracking approach can be found in literature. For example the authors in [53] propose an on-line learning algorithm to learn a discriminative appearance model in a multi-target tracking method using an AdaBoost algorithm. The authors use the defined model for each target to compute the association between targets and new observations. For each tracklet they compute an affinity score evaluating the statistic information in a sliding window, in order to consider a certain interval time.

The authors in [94] introduce an online learning approach for multi-target tracking considering a Conditional Random Field (CRF) model. They define an energy function to manage the tracking process that consider motion and appearance models of each tracked target. They also consider a global model used with a sliding window process to minimize problems like fragmentation or identity switch.

### 2.1.3 Tracking-by-detection

An important class of algorithms used for visual tracking are the *tracking-by-detection* approaches, where detections are collected frame-by-frame to reconstruct trajectories. In [66] the authors propose a technique that uses a deformable part models in a Kalman filtering framework. The main idea is that a pedestrian consists of multiple parts, and the detection system should localize these multiple parts and maintain only those detections with some parts next to each other. To overcome the problem of missed detection they use the Kalman filter to estimate the state of each target. On the other hand, the data association problem is solved by mean of measurements based on appearance features as well as geometry.

In [84] the authors propose a part-based person-specific SVM classifiers which capture the articulations of dynamically changing human bodies and background. In this way they are able to handle partial occlusions in both the detection and the tracking stages. In the detection stage, they select the subset of parts which maximizes the probability of detection, and significantly improves the detection performance in crowded scenes. To increment the discriminative power of the classifier, the authors use as positive example the detection from all detections of the trajectories, and similar negative examples from the other trajectories. To solve the data association problem the authors use a first-order Markov model in which trajectories are continuously growing as the tracking proceeds. Detections are associated with existing trajectories by a greedy bipartite algorithm based on an affinity matrix between trajectories and detections.

Another type of tracking-by-detection approach based on Markov model is presented in [80]. In this work the authors consider two types of particle filter: a local and a global one. They are integrated in a Markov framework, using the local particle filter to handle the association process between tracklets and observations, and the global particle filter which are used to capture the interaction between the objects. The Markov model manages the interaction between the two different particle filters, to define an approach robust to missed detection or occlusion.

The authors in [81] present a part-based model tracking method in combination with a particle filter. They use particle filters to estimate each part location to speed up the scanning area process and to determine the new observations. The association process is performed considering a score based on the quality of the detection for a certain tracklet. This score considers the estimation process allowing the improvement of the association between tracklets and new observation.

As proposed in other works [92, 94], in [13] the authors define a technique for tracking process that considers a global data association process to decrease common problems in a tracking system. They define a hierarchical grid-based people detector, that produces an accurate localization on the ground plane, in order to increase the targets detection and the frame-by-frame association process. They also define a global data association process in order to enable the tracker to recover miss detections or occlusions for a long-term interaction between the targets during the tracking process.

Another case with a tracking algorithm that combine local and global information is proposed in [102], where the authors define a unified algorithm to find the joint optimal association. This algorithm learns a trajectory model for each target, based on information like appearance model, velocity, size and position. An iterative process updates each trajectory model to improve the quality of the tracklet and the efficiency of the approach.

The state-of-the-art of the facial characteristic estimation is presented in the following section.

## 2.2 Facial Characteristic Estimation

Most systems for estimating target characteristics like Age, Gender [54, 103] and head pose [34, 44] have been proposed and developed as isolated issues. Each type of estimator uses its own sets of custom features and its own specific estimation techniques. In this section we present some of the recent literature on face characteristic estimation from images.

### 2.2.1 Gender Classification

Among the facial characteristics, Gender is the most commonly studied and there is a vast literature on Gender classifiers. In general a Gender classifier is based on the creation of a custom feature, and the use of an SVM classification approach. In [54] the authors recognize Gender using Local Binary Pattern (LBP) features extracted on small facial regions and using a support vector machine (SVM) to perform the classification. The authors in [91] propose a Gender classifier based on the use of local Gabor binary mapping pattern (LGBMP), obtained mapping the local Gabor binary pattern (LGBP) in a low-dimensional feature, considering an SVM classifier and tested it with a public dataset. Guo *et al.* [37] demonstrated that Gender recognition based on LBP and the Histogram of Oriented Gradients (HOG) [20] features can be helped by Age classification.

Zheng *et al.* [103] propose a support vector machine with automatic confidence (SVMAC) for pattern classification. They develop an algorithm incorporating the label confidence of each training sample into the learning process and use quadratic programming to solve the optimization. Moreover, they use feature selection methods that extract the subregions from the images and obtain the best results using LGBPs. See [68] for a detailed survey on Gender estimation.

### 2.2.2 Age Estimation

It must be observed that the interest on the Age estimation significantly increased in the last years. There is a vast literature on Age estimation alone or jointly treated with other particular facial characteristic. Gunay *et al.* [35] propose an Age classification approach based on a nearest neighbour algorithm with LBP feature, obtaining about $80\%$ accuracy on a subset of the FERET dataset [72].

In [75] the authors introduce the concept of facial aging to improve the efficiency of Gender estimation and face recognition and proposing the “MORPH” dataset [77] as a benchmark for Gender and Age estimation. The authors in [76] propose a feature extraction based on structural face characteristic, considering an edge detector and a neural network process, while in [99] propose an approach to Age classification that uses the LBP feature.

Ma *et al.* [60] propose an Age estimation method based on a double layer multiple task learning approach which applies SVM methodologies. The proposed approach considers a local model for each subject and a global model used to share knowledge of common aging pattern among different local models. In the experimental test they consider a combination between a Gabor filter and Principal Component Analysis (PCA) approach for the feature extraction phase.

### 2.2.3 Soft Biometric Estimation

One of the first works aimed to study a method to jointly estimate multiple biometric characteristics is presented in [96]. The authors use AdaBoost on LBPH features to jointly estimate Gender, Age and Ethnicity. In particular in [52] Kumar *et al.* propose a technique for face verification in uncontrolled environments, considering two different approaches: one based on attribute classifiers for Gender, Age and Ethnicity, and a second one to recognize facial landmarks. The main goal of this work was face verification in the wild, and they tested it on two datasets: LFW [45] and PubFig [52]. Considering the dataset dimensions, around 13,000 of 5,749 people and 60,000 images of 500 people, respectively, they achieved excellent results with about 80% of recognition accuracy.

With the same goal, Guo *et al.* in [36] investigate two different ways to estimate Gender, Age and Ethnicity using Kernel Canonical Correlation Analysis (KCCA) and PLS (Partial Least Squares). They show that it is possible to jointly estimate Age, Gender and Ethnicity, while feature dimensionality is significantly reduced.

Considering only Age and Gender characteristics, the authors in [31] and the authors in [24] propose two different dataset, *Groups* and *Adience* respectively, using a set of images extracted from Flickr, considering common situations. In [31] the authors propose a contextual feature extraction with a Gaussian Maximum Likelihood classifier to define Age and Gender. Eidinger *et al.* [24] propose a SVM classifier to estimate Age and Gender in the wild. In this case they consider different type of feature to discriminate properly each characteristic.

### 2.2.4 Head Pose Estimation

Head pose, while not a soft biometric, is a fundamental characteristic in video surveillance as it can provide an estimate of interest for profiling-at-a-distance applications [49], or evidence of the direction in which a person is moving [7]. In [87] the authors use a neural network approach to estimate a monocular head pose estimation using a classifier to segment skin color cluster, and a *multi-view head pose estimation* to define the head pose in a real scenario, with the goal of studying human activities and interactions to provide helpful services implicitly and unobtrusively.

The authors of [86] use an approach based on tensors that localize the nose-tip of a face in a coarse-to-fine fashion. On the other hand in [59] the authors use Local Gabor Binary Patterns (LGBPs) to encode the local facial characteristics, keeping the spatial information and considering many subregions of the image. In [42] the authors propose a technique for head pose estimation that use the extraction of Scale-invariant feature transform (SIFT) with a SVM classification approach. They use a Random Projection to reduce the dimensionality of a dense SIFT, and a combination between a SVM classifier and Support Vector Regression (SVR) to define their classifier.

In [7] the authors propose an unsupervised method to estimate gaze direction with random decision trees over Histogram of Oriented Gradients (HOG) and color features. In [38] partial least squares (PLS) and a 3-level pyramid of HOGs are proposed to estimate head pose in the presence of misalignment.

# Chapter 3

# Data Association using $\ell_1$-regularization

*"Never forget what you are, for surely the world will not. Make it your strength. Then it can never be your weakness. Armour yourself in it, and it will never be used to hurt you."*

- George R.R. Martin, *A Game of Thrones.*

Computer vision applied to video surveillance applications like abnormal behaviour detection, group interaction analysis, and object tracking has received a lot of attention in the last decade. One of the most important tasks related to these topics is visual object tracking. The task of multiple target tracking is to follow targets in an uncontrolled environment handling at the same time problems such as occlusion, similarity in the target appearance and crowded scenes.

The data association (DA) problem is one of the main hurdles to overcome in multiple target tracking and consists of finding the right assignment between the set of tracked targets and the set of new observations extracted from the current frame of a sequence. For each tracked target, the past observations that have already been associated with it compose a tracklet (in Figure 3.3 three different example of tracklets are shown). In Figure 3.1 on the left are shown three tracklets corresponding to three tracked subjects and on the right are three new observations that must be associated to these targets. This task may become difficult in real-world scenarios due

The work presented in this chapter has been published as "Multi-Target Data association using sparse Reconstruction" in *Proceedings of International Conference on Image Analysis and Processing (ICIAP), September, Naples, Italy, 2013 [4].*

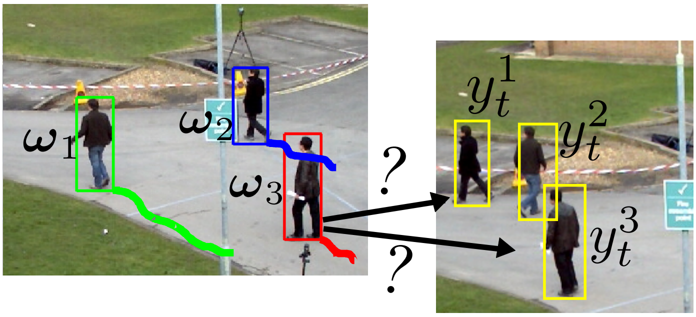


Figure 3.1: The data association problem. At each time instant observations $y_t^i$ must be associated with one of the running trackers $\omega_k$.

to many problems that may arise. One problem is how to create a representation that discriminatively models each target through time, while another one is how to build an accurate rule for discerning each subject from the others in the scene. Moreover, data association must scale well with the number of targets considering real time constraints.

For example with the Bayesian data association approach [16, 51, 61] for each target a probabilistic model defined with a state function is considered. In particular each state can be composed by the position observed and the velocity, using these information to produce hypotheses that associate tracklets with new observations. In [2] the authors define a Kalman based model, that uses an energy function to define a global model that, for each target, evaluate all the associated observation within a temporal window, representing the existence, motion and interaction of all the objects.

In addition to the model representation, we must take into account several issues such as target occlusion and ID switches introduced in Chapter 2, that usually complicate the data association problem for many targets. In those cases we observe that approximations are often applied. In this chapter we introduce a solution to multi-target data association problem based on $\ell_1$-regularized sparse basis expansions. Assuming sufficient training samples per subject are available, our idea is to create a discriminative basis of observations that can be used to reconstruct and associate a new target. The use of $\ell_1$-regularized basis expansions allows our approach to exploit multiple instances of the target when performing data association rather than relying on an average representation of target appearance. This usually allows to define the most likely assignment between tracked targets and new observations. The experimental results on the PETS dataset [1] are encouraging and demonstrate that our method is sufficiently accurate and efficient for multi-target data association.

In the next section we briefly review some works based on sparse methods relevant to computer vision application and used in the present research activity. In Section 3.2 we describe a technique that uses sparse basis expansion to solve the multi-target data association problem. Finally, we report some preliminary experiments in Section 3.3 and discuss ongoing work in Section 3.4.

## 3.1 Sparse Reconstruction and Sparse Coding

*Sparse methods* are often used in computer vision community [48, 57, 58, 74, 95, 100]. These approaches exploit the hypothesis that an arbitrary signal can be reconstructed using a sparse combination of (potentially many) basis vectors. These approaches are based on the resolution of an optimization problem that uses a certain type of regularization, often based on $\ell_1$ or $\ell_2$ norm. They have been applied to tracking [65], face recognition [88] and vocabulary construction [62], where the target model is usually defined considering a certain set of observations that is used in an optimization problem which we present in the following. For example, in [65] a model for a single tracker is defined: it uses the acquired information adding artificial data to handle the partial occlusions in the association process. The authors in [88] propose a global model to represent all subjects with a dictionary of gallery faces.

It must be observed that the usage of a regularization model rather than another, changes the way the proposed approach is applied. The more common formulation used in computer vision is known in the statistics literature as Lasso Regression [41, 85], that uses an $\ell_1$ norm to compute the regularization. The optimization problem is defined as:

$$\min_{\boldsymbol{\alpha}} \|\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1, \tag{3.1}$$

where:

- $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_D]$ is the matrix of basis vectors $\mathbf{b}_d$ which are used for reconstructing a new signal.
- $\mathbf{y}$ is the new signal that we want to reconstruct.
- $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D)'$ represents the vector of reconstruction coefficients for each basis vector.
- $\lambda$ controls the trade off between minimization of the $\ell_2$ reconstruction error and the $\ell_1$ norm of the coefficients used to reconstruct $\mathbf{y}$.

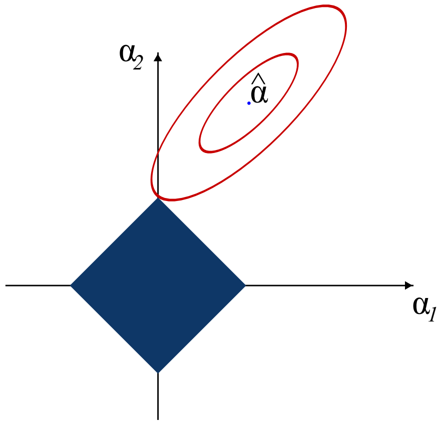


Figure 3.2: Graphical exemplification of the optimization process with $\ell_1$ regularization for the reconstruction of a bi-dimensional case (Eq. (3.1)) [41, 85].

Regularized basis expansions of this type are generally referred to as *sparse* because the $\ell_1$ regularization term, depending on the sparseness factor $\lambda$, tends to cause the coefficients of reconstruction to collapse to zero except for a few important basis vectors. In Figure 3.2 a bi-dimensional case ($D = 2$) of the sparse reconstruction is shown. The optimization method finds a solution $\hat{\boldsymbol{\alpha}}$ considering the condition $|\alpha_1| + |\alpha_2| \leq \lambda$, to guarantee the sparseness of the $\hat{\boldsymbol{\alpha}}$ coefficients. This particular aspect makes the *Sparse coding* a more generalist approach. It is used for compression audio purposes [73], for improving the performance on web image ranking in a text-based image search [100] and in other computer vision applications.

In visual tracking, sparse reconstruction has been applied to the single-target tracking problem [65]. This approach tries to find the best association between target and observation using a basis composed of past target observations. With a similar purpose the authors in [46] define a single target tracking approach proposing a local sparse model robust to occlusion. The tracking process models a target with a set of overlapped patches using it with their sparse association approach. The proposed model is robust to partial occlusion due to the redundancy of the target model, thanks to the fact that each patch can contain target information and spatial information, locating more precisely the tracked target. Moreover, the model is updated online on the basis of an incremental subspace learning and sparse representation.

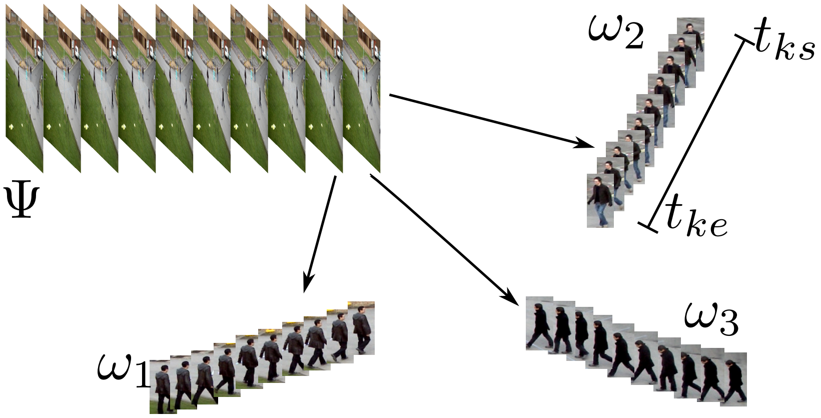


Figure 3.3: An exemplification of tracklets creation is shown in figure. From a video $\Psi$ a data association method creates three different tracklets $\omega_k$.

To define a multi-target algorithm approach in [90] a method that solves the detection and the data association problems jointly is proposed. It considers a sparse approach to detect new observations and a first-order Markov chain to perform the data association phase. In a discriminative classification setting, sparse reconstruction has also been applied to face recognition problems [88]. In the above mentioned work the authors define an approach to face recognition that uses sparse reconstruction of probe images in terms of a dictionary of gallery faces. Classification of an unknown face is performed using the reconstruction error of sparse basis expansions. In other cases sparse methods are used to define an appearance model as in [104].

## 3.2 Data Association by Sparse Reconstruction

This method is focused only on *pure data association* problem, assuming perfect detections and perfect bootstrapping of appearance models in order to isolate data association performance from the complexities of multi-target tracking. In this section we explain and analyze each stage of the proposed method.

### 3.2.1 The Data Association Problem

In general, a data association problem is the association of measurements with models (trackers, in this case) at each time step of a sequence. More formally, considering a video stream $\Psi$ whose duration is $T \in \mathbb{N}$ seconds, suppose that $K$ different targets moving in the scene can be identified. Now consider a particular target $k$ observable in the time interval $[t_{ks}, t_{ke}] \subset [1, T]$, where $t_{ks}$ is time of the first appearance and $t_{ke}$ is the last appearance or exit time (hence $t_{ks} < t_{ke}$) (see Figure 3.3). For each time instant $t$, we consider a perfect detector that allows to obtain a set of observations $y_t$ with a cardinality $L_t \in \mathbb{N}$, such that:

$$\begin{aligned} Y &= \{y_t : t \in [1, T]\}, \\ y_t &= \{y_t^i\}_{i=1}^{L_t}. \end{aligned} \tag{3.2}$$

For each instant $t$ the value of $L_t$ may be different.

A tracking algorithm is aimed to define a set of tracklets:

$$\Omega = \{\omega_k : k \in [1, K]\}. \tag{3.3}$$

Considering the set $Y$, each tracklet $\omega_k$ will be characterized by a sub-set of observations, where each observation of $\omega_k$ belongs to a distinct time instant:

$$\omega_k = \{y_t^l : l \in [1, L_t], \forall t \in [t_{ks}, t_{ke}]\} \subseteq Y. \tag{3.4}$$

Note that an observation $y_t^i$ can only be associated with a single tracklet $\omega_k$:

$$\omega_k \cap \omega_j = \emptyset, \forall k, j \in [1, K] \text{ if } k \neq j. \tag{3.5}$$

In the following, $k$ will refer to a target. A good tracker needs a good data association method to correctly associate observations $y_t^i$ and tracklets $\omega_k$.

### 3.2.2 Sparse Discriminative Basis Expansion for Data Association

In this section we will first introduce the used feature descriptor and then we will explain how we build discriminative bases for data association. Using a regularized sparse basis expansion we have been able to compute a reconstruction error according to the basis of each existing target. Finally, we present how our algorithm solves the data association problem by combining the reconstruction error with spatial proximity information.

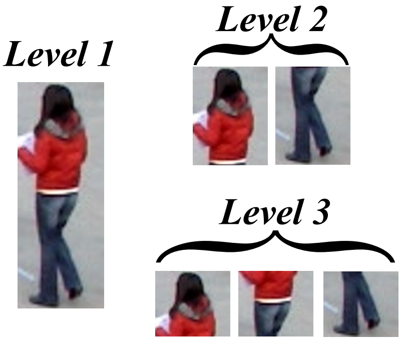


Figure 3.4: The three level pyramid used in the feature representation are represented.

### Feature Adopted

For each observation $y_t^i$ we extract a feature vector $\mathbf{f}(y_t^i)$. In this approach, a pyramidal color histogram is used in order to obtain a multi-level representation of the appearance of each detection. We define a three level pyramid, where the top level corresponds to the full detection window, the second level to two non-overlapping horizontal slices and the third and last level to three horizontal slices (see Figure 3.4).

Each slice is represented by a RGB color histogram $\mathbf{h}_i$ which is normalized with the $\ell_1$ norm, while the whole feature vector $\mathbf{h}$ is normalized with the $\ell_2$ norm. This feature maintains multi-level appearance information, that is defined as a vector $\mathbf{f}(y_t^i) \in \mathbb{R}^m$ with $m = |\mathbf{h}| = 3072$ bins. It must be observed that it does not rely on complex foreground/background segmentation or part models, and that it has good illumination invariance and good independence with respect to the quality of the observations.

### Discriminative Basis Construction

The key idea behind the proposed approach is the construction and the usage of a discriminative basis $\mathbf{B}$ that, when used to perform a sparse reconstruction of an unknown target, can be exploited to recover which basis vectors of a tracked target mostly contribute to the sparse reconstruction. Assuming that $n$ observations have already been associated with the $k$-th tracked target, we define the sub-basis corresponding to target $k$ as the concatenation of the $n$ feature descriptors of all

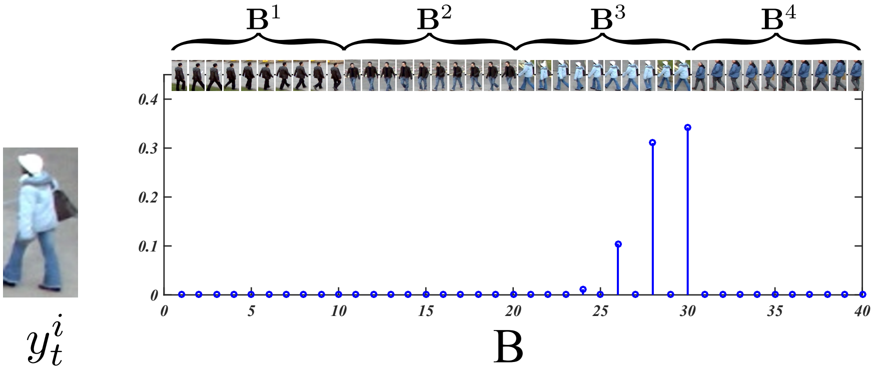


Figure 3.5: Sparse basis expansion. A new observation $y_t^i$ is reconstructed in terms of a basis of tracklet observations. The $x$-axes indexes the coefficients $\alpha_k^l$ of $\boldsymbol{\alpha}$ in the reconstruction computed using Equation (3.7). Note how the reconstruction is *sparse* in that only a few basis vectors have non-zero $\alpha_k^l$ coefficients.

associated observations:

$$\mathbf{B}^k = \left[ \mathbf{f}(y_{t_1}^{i_1}), \ldots, \mathbf{f}(y_t^i), \ldots, \mathbf{f}(y_{t_n}^{i_n}) \right], \;\; \forall\, y_t^{i_t} \in \omega_k. \tag{3.6}$$

If there are $K$ targets in the scene, the discriminative basis $\mathbf{B}$ is obtained by concatenating these sub-basis, which is hence composed of $D = K \cdot n$ feature vectors, i.e. $D = |\mathbf{B}|$. As noted at the beginning of the section, we consider a perfect bootstrapping, inspired by the concept of *reliable tracklet* in [53] with a bootstrapping of $n = 25$ frames for each target.

### Regularized Sparse Basis Expansion

The upper part of Figure 3.5 depicts an example of a discriminative basis. In this case the basis $\mathbf{B}$ is called *discriminative* because each vector $\mathbf{b}_l^k \in \mathbf{B}$ is a feature vector $\mathbf{f}(y_t^l)$ associated with a specific target label (i.e. we maintain discriminative information about each sub-basis). Solving an $\ell_1$-regularized optimization problem:

$$\min_{\boldsymbol{\alpha}} \|\mathbf{f}(y_t^i) - \mathbf{B}\boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1, \tag{3.7}$$

allows to obtain a sparse projection vector $\boldsymbol{\alpha}$. The $\boldsymbol{\alpha}$ vector is composed of $D$ coefficients that indicate how to reconstruct a new observation $\mathbf{f}(y_t^i)$ using a linear combination of the sample vectors in $\mathbf{B}$. The coefficient $\lambda \in \mathbb{R}^+$ in Equation (3.7)

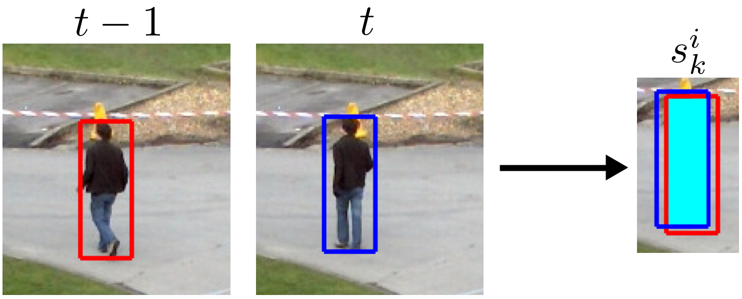


Figure 3.6: Exemplification of the VOC score used to compute the spatial proximity information. Considering the bounding box area of the same target in two different instants, the score $s_k^i$ quantifies the colored area.

is used to control the sparsity of $\boldsymbol{\alpha}$: the larger the value of $\lambda$, the lower the $\ell_1$ norm of the projection vector $\boldsymbol{\alpha}$. In the lower part of Figure 3.5 the magnitudes of the reconstruction coefficients $\alpha_k^l$ are depicted for an observation sample.

To estimate which target $k$ should be associated with a new observation $\mathbf{f}(y_t^i)$, we analyze the reconstruction coefficients in $\boldsymbol{\alpha}$. The vector $\boldsymbol{\alpha}$ can be seen as a concatenation of $\boldsymbol{\alpha}_k$ that are the coefficients corresponding to each target, such that

$$\boldsymbol{\alpha} = \begin{bmatrix} \boldsymbol{\alpha}_1 & \boldsymbol{\alpha}_2 & \cdots & \boldsymbol{\alpha}_k & \cdots & \boldsymbol{\alpha}_K \end{bmatrix}', \tag{3.8}$$

where each $\boldsymbol{\alpha}_k$ is composed of $n$ different $\alpha_k^l$, one for each basis template $\mathbf{b}_l^k$ of $\mathbf{B}$. To identify the associated targets we define a *reconstruction error* $\varepsilon_k^i$ for each $(k, i)$:

$$\varepsilon_k^i = \|\mathbf{f}(y_t^i) - \mathbf{B}^k \boldsymbol{\alpha}_k\|_2. \tag{3.9}$$

The value $\varepsilon_k^i$ corresponds to the reconstruction error when $\mathbf{f}(y_t^i)$ is reconstructed using only those coefficients from $\boldsymbol{\alpha}$ and columns from $\mathbf{B}$ that correspond to tracked target $k$. Since the feature is $\ell_2$ normalized, $\varepsilon_k^i \in [0, 1]$. The smaller the error $\varepsilon_k^i$, the greater the likelihood that $y_t^i$ represents the target $k$.

### Spatial Proximity Information

The VOC Score [26] between trackers and new observations are being used to combine spatial proximity and sparse reconstruction error. The spatial proximity score between tracker $k$ and observation $i$ is:

$$s_k^i = \frac{A_k \bigcap A_i}{A_k \bigcup A_i}, s_k^i \in [0, 1], \tag{3.10}$$

**Algorithm 1:** Data association algorithm

```
**Data**: $\mathbf{B}$, $\Omega$, $y_t$ and $\gamma$
$\Omega_t = \Omega$ : local set of tracklets ;
compute $\mathbf{f}(y_t^i)\,\forall y_t^i$ ; $s_k^i, \varepsilon_k^i, a_k^i\,\forall i, \forall k$ ;
**while** $\Omega_t \neq \emptyset \wedge y_t \neq \emptyset$ **do**
    $(\hat{k}, \hat{i}) = \arg\min_{k,i} a_k^i$;
    $\omega_{\hat{k}} = \omega_{\hat{k}} \cup \{y_t^{\hat{i}}\}$ ;
    $y_t = \{y_t \setminus y_t^{\hat{i}}\}$ ;
    $\Omega_t = \{\Omega_t \setminus \omega_{t,\hat{k}}\}$ ;
**end while**
```

where $A_k$ is the bounding box area of the potential last observation $y_\tau^l$ associated with the tracklet $\omega_k$, and $A_i$ is the area of the new observation $y_t^i$.

The VOC score corresponds to the overlap of $A_k$ and $A_i$ normalized by the union of the areas. An example of the overlapped area used to compute the VOC score is shown in Figure 3.6. If the areas are highly overlapping the score $s_k^i$ will tend to one, while its value will tend to zero if the overlap is small. Note that $\tau \in [t-5, t-1]$ i.e. we only compute $s_k^i$, if there is an association with tracker $\omega_k$ in the last 5 frames.

Finally, the VOC Score is used in combination with the reconstruction error $\varepsilon_k^i$ introduced in the previous section to define the association error $a_k^i$:

$$a_k^i = (1-\gamma)\varepsilon_k^i + \gamma(1 - s_k^i), \;\; \forall (k,i) \in [1, K] \times [1, L_t]. \tag{3.11}$$

The parameter $\gamma$ is used to control the trade-off between spatial and sparse reconstruction in determining the association error.

### 3.2.3 Data Association Algorithm

In this section the concepts presented in the previous sections are considered jointly in order to define an algorithm for data association using sparse basis expansions. The purpose of this preliminary study is to determine the potential of sparse methods for data association under ideal conditions, making a number of simplifying assumptions. It is important to consider that, we assume that perfect detections are available for all people appearing in the video stream and that the first $n$ observations of each target can be perfectly associated. This effectively allows us to create the discriminative basis $\mathbf{B}$, and to update it when new targets appear in the scene.

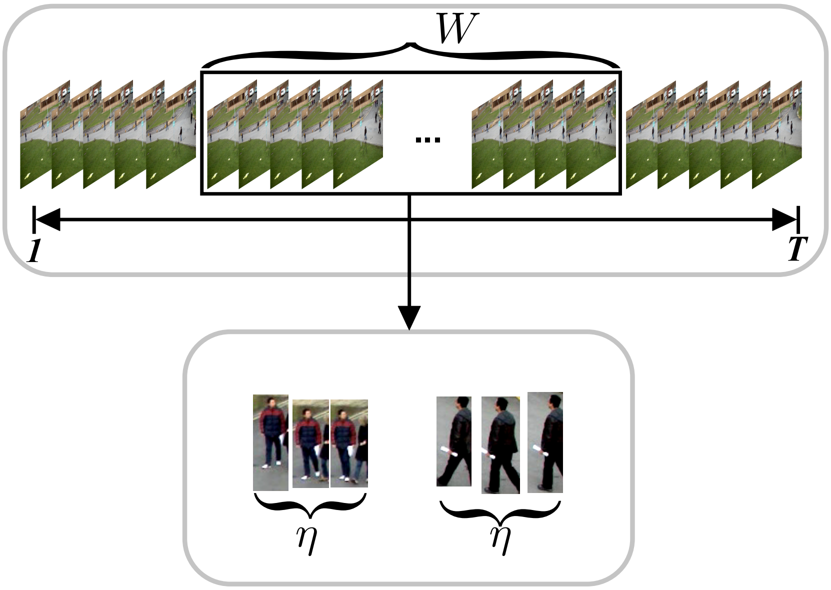


Figure 3.7: Update process. Considering a temporal window $W$, for each sub-basis is being extracted at most $\eta$ new detection to update the model.

At each time instant $t$, we compute for each tracklet/observation pair the reconstruction error and spatial proximity that are combined to form the association error, see Equation (3.11). In this way a new observation is associated with one of the existing tracklets according to the greedy Algorithm 1.

During the tracking process the information contained in the discriminative basis may become outdated and thus may no longer describe well a particular target $k$. In our data association algorithm we include an update phase that adds a fixed number of feature vectors for each target (see Figure 3.7). Basis update is performed by exploiting the associations occurring in a temporal window of $W$ frames. For each tracklet we add at most the $\eta$ best associated observations (according to Equation (3.11)) to the corresponding sub-basis.

With the update phase, the discriminative basis size may increase and be different for each person after the update, i.e. we may have $n_k \neq n \, \forall k$. However we observe that the $\ell_1$-regularization in our reconstruction will always tend to give a sparse projection vector. In this way only few elements of each sub-basis are used in the presented reconstruction process.

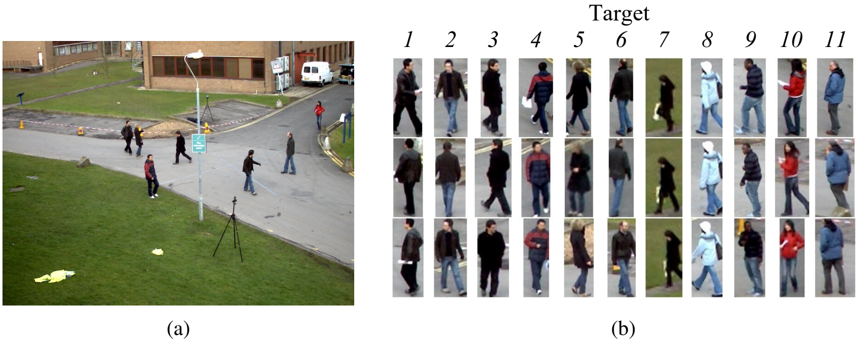


Figure 3.8: PETS dataset. (a) A sample frame of the dataset. (b) Three instances of each tracked subject in the sequence.

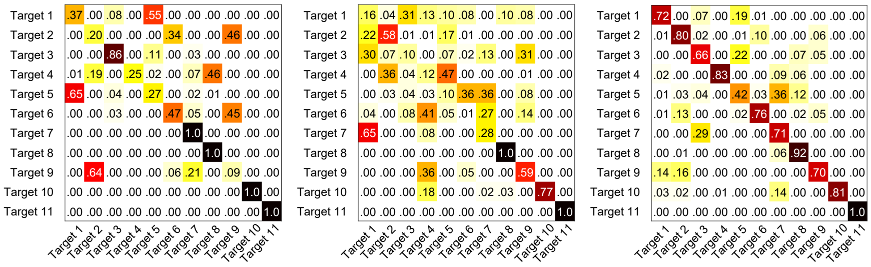
## 3.3 Experiments

In this section we report about some experimental results obtained with our method, varying the parameters $\lambda$, $\gamma$ with or without the basis update phase. Experiments are performed on the *s2.l1-view01* sequence of the PETS 2009 public dataset [1] (see Figure 3.8(a)), that is a de-facto standard in the tracking community due to its challenging nature. It is one of the most used sequence in the literature on multi-target tracking [11, 93]. Comparison with state-of-the-art methods are obtained using CLEAR MOT [8] metric.

### 3.3.1 Data Association Performance

The results obtained by varying $\lambda$ and $\gamma$ are shown as confusion matrices in Figure 3.9. These matrices give a clear idea of the potential of the method for pure multi-target data association under ideal tracking conditions. The matrix shown in Figure 3.9(a) was obtained with $\lambda = 0.7$, $\gamma = 0.5$ and without updating the discriminative basis. This value of $\lambda$ means that the behavior of data association is similar to the NNSF. From this matrix, we can observe that nearest neighbor approaches can result in many association errors, in particular for targets 1, 2, 4, 5 and 6. This is principally due to the fact that these targets have similar appearance and frequently occlude each other.

The matrix shown in Figure 3.9(b) was obtained with $\lambda = 0.1$ and $\gamma = 1$, without updating the discriminative basis. This value of $\gamma$ forces the proposed



(a)

| | Target 1 | Target 2 | Target 3 | Target 4 | Target 5 | Target 6 | Target 7 | Target 8 | Target 9 | Target 10 | Target 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Target 1 | .37 | .00 | .08 | .00 | .55 | .00 | .00 | .00 | .00 | .00 | .00 |
| Target 2 | .00 | .20 | .00 | .00 | .00 | .34 | .00 | .00 | .46 | .00 | .00 |
| Target 3 | .00 | .00 | .86 | .00 | .11 | .00 | .03 | .00 | .00 | .00 | .00 |
| Target 4 | .01 | .19 | .00 | .25 | .02 | .00 | .07 | .46 | .00 | .00 | .00 |
| Target 5 | .65 | .00 | .04 | .00 | .27 | .00 | .02 | .01 | .00 | .00 | .00 |
| Target 6 | .00 | .00 | .03 | .00 | .00 | .47 | .05 | .00 | .45 | .00 | .00 |
| Target 7 | .00 | .00 | .00 | .00 | .00 | .00 | 1.0 | .00 | .00 | .00 | .00 |
| Target 8 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | 1.0 | .00 | .00 | .00 |
| Target 9 | .00 | .64 | .00 | .00 | .00 | .06 | .21 | .00 | .09 | .00 | .00 |
| Target 10 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | 1.0 | .00 |
| Target 11 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | 1.0 |

(b)

| | Target 1 | Target 2 | Target 3 | Target 4 | Target 5 | Target 6 | Target 7 | Target 8 | Target 9 | Target 10 | Target 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Target 1 | .16 | .04 | .31 | .13 | .10 | .08 | .00 | .10 | .08 | .00 | .00 |
| Target 2 | .22 | .58 | .01 | .01 | .17 | .01 | .00 | .00 | .00 | .00 | .00 |
| Target 3 | .30 | .07 | .10 | .00 | .07 | .02 | .13 | .00 | .31 | .00 | .00 |
| Target 4 | .00 | .36 | .04 | .12 | .47 | .00 | .00 | .00 | .01 | .00 | .00 |
| Target 5 | .00 | .03 | .04 | .03 | .10 | .36 | .36 | .00 | .08 | .00 | .00 |
| Target 6 | .04 | .00 | .08 | .41 | .05 | .01 | .27 | .00 | .14 | .00 | .00 |
| Target 7 | .65 | .00 | .00 | .08 | .00 | .00 | .28 | .00 | .00 | .00 | .00 |
| Target 8 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | 1.0 | .00 | .00 | .00 |
| Target 9 | .00 | .00 | .00 | .36 | .00 | .05 | .00 | .00 | .59 | .00 | .00 |
| Target 10 | .00 | .00 | .00 | .18 | .00 | .00 | .02 | .03 | .00 | .77 | .00 |
| Target 11 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | 1.0 |

(c)

| | Target 1 | Target 2 | Target 3 | Target 4 | Target 5 | Target 6 | Target 7 | Target 8 | Target 9 | Target 10 | Target 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Target 1 | .72 | .00 | .07 | .00 | .19 | .01 | .00 | .00 | .00 | .00 | .00 |
| Target 2 | .01 | .80 | .02 | .00 | .01 | .10 | .00 | .00 | .06 | .00 | .00 |
| Target 3 | .00 | .00 | .66 | .00 | .22 | .00 | .00 | .07 | .05 | .00 | .00 |
| Target 4 | .02 | .00 | .00 | .83 | .00 | .00 | .09 | .06 | .00 | .00 | .00 |
| Target 5 | .01 | .03 | .04 | .00 | .42 | .03 | .36 | .12 | .00 | .00 | .00 |
| Target 6 | .01 | .13 | .00 | .00 | .02 | .76 | .00 | .02 | .05 | .00 | .00 |
| Target 7 | .00 | .00 | .29 | .00 | .00 | .00 | .71 | .00 | .00 | .00 | .00 |
| Target 8 | .00 | .01 | .00 | .00 | .00 | .00 | .06 | .92 | .00 | .00 | .00 |
| Target 9 | .14 | .16 | .00 | .00 | .00 | .00 | .00 | .00 | .70 | .00 | .00 |
| Target 10 | .03 | .02 | .00 | .01 | .00 | .00 | .14 | .00 | .00 | .81 | .00 |
| Target 11 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | 1.0 |

Figure 3.9: Confusion matrices for various parameter settings. (a) with $\lambda = 0.7$, $\gamma = 0.5$, no update phase. (b) spatial proximity only with $\gamma = 1$. (c) with $\gamma = 0.2$ and $\lambda = 0.1$, basis update with $W = 20$ and $\eta = 3$.

association to give more importance to the spatial proximity score. However, as shown in the confusion matrix, the association accuracy is low for the majority of the targets.

On the other hand the matrix of Figure 3.9(c) is obtained by using $\lambda = 0.1$ and $\gamma = 0.2$, with the update of the discriminative basis with $W = 20$ and $\eta = 3$. In this confusion matrix we can observe that this configuration gives the best association results for each target with respect to the other configurations. From these results we can conclude that values of $\lambda$ which enforce sparsity without reducing the projection vector $\boldsymbol{\alpha}$ to a single non-null value are preferable and that results are improved when a contribution from spatial proximity is added.

### 3.3.2 Comparison with the State-of-the-art

In this section we discuss the results obtained with the CLEAR MOT metrics [8]. In particular, Table 3.1 reports results with three different configurations of proposed method corresponding to $\gamma \in \{0.2, 0.4, 0.8\}$ and some results from state-of-the-art techniques on this sequence [11, 93]. It must be noted that these results are only a byproduct of data-association process and therefore they are not directly comparable with the state-of-the-art methods. This is due to the fact that we make many simplifying assumptions about perfect detections and initial discriminative basis construction. This comparison is only intended to give some indication of the potential of our data association, considering that we focused only on the *pure data association problem.*

| Method | MOTA | Recall | Precision | FN Rate | FP Rate | IDS |
|---|---|---|---|---|---|---|
| Yang [93] PM Only | – | 92.8% | 95.4% | – | – | 0 |
| Yang [93] PM + CFT | – | 97.8% | 94.8% | – | – | 0 |
| Breitenstein et al. [11] | 79.7% | – | – | – | – | – |
| Our $\ell_1$-DA ($\gamma = 0.2$) | 82.8% | 82.9% | 96.2% | 13.9% | 0.04% | 146 |
| Our $\ell_1$-DA ($\gamma = 0.4$) | 84.7% | 84.8% | 98.4% | 13.9% | 0.02% | 60 |
| Our $\ell_1$-DA ($\gamma = 0.8$) | 80.5% | 80.5% | 99.9% | 19.4% | 0% | 4 |

Table 3.1: Results on the *s2.l1-view01* sequence of the PETS 2009 dataset.

For these experiments we set $\lambda = 0.1$ and update the basis with $W = 20$ and $\eta = 3$. In Table 3.1 it can be observed that varying $\gamma$ mostly affects the number of identity switches (IDS) and the accuracy (MOTA) of the solution. MOTA indicates the accuracy of the approach in terms of multi-target tracking. The recall is computed as the total number of true positives over the total number of ground truth objects, while the precision is calculated as the total number of true positive over the sum of the number of active tracks over frames. The switch of target identities affects the precision because one id switch induces one less true positive. Moreover, the proposed solution does not deal with target that are no longer present in the scene because they are not removed from the model. This can result in a false negative if a detection is associated to a target that has exited the scene.

From Table 3.1 it is possible to observe that, with a low value of $\gamma$, the technique gives more importance to the appearance and this can result in a identity swaps since very little spatial proximity information is considered. However, by increasing the contribution of spatial proximity score in the association score we observe a dramatic decrease in identity switches, though this can also result in lower accuracy. From these results, setting $\gamma = 0.4$ seems to be a good trade-off between MOTA and IDS. Our method performs competitively with the state-of-the-art, even compared to offline tracking methods that require all detections beforehand to perform association and extract trajectories of each target [93].

## 3.4 Discussion and Conclusions

In this chapter we propose an approach to multi-target data association that exploits sparse reconstruction and spatial proximity. We show that integrating multiple templates of each target in a discriminative basis helps in the association process, but at the same time spatial constraints were required to obtain good performance for the tracking of multiple targets. With these preliminary results we show that, under very strict ideal detection and initialization hypotheses, our data association method can be competitive with those used by state-of-the-art tracking methods.

In the next chapter the propose data association algorithm was verified in a realistic situation, defining an end-to-end tracking-by-detection approach able to improve the long-term consistency.

# Chapter 4

# Multi-Target Tracking using the Weighted Sparse Reconstruction

*It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.*

- Arthur Conan Doyle, *A Scandal in Bohemia.*

In this chapter we describe an approach to the multi-target tracking problem based on $\ell_1$-regularized sparse basis expansions. Our method is based on the creation of a discriminative basis of observations that we used to reconstruct and associate new targets in the data association phase. The use of $\ell_1$-regularized basis expansions allows our method to exploit multiple instances of the target when performing data association rather than relying on an average representation of target appearance. Moreover, we propose a new global association phase with the goal to reduce the number of ID switches and improve the overall performance. Experimental results show that the proposed approach yields comparable performance with respect to the state-of-the-art on both the PETS and ETH datasets, while also reducing the number of ID switches.

Multi-target visual tracking is the robust and accurate prediction of target locations in video streams. It is a fundamental component in many video surveillance

Part of this chapter has been submitted to *Pattern Analysis and Application* and under review, October, 2015.

systems which must identify and localize people, cars, and other moving objects in real time. Depending on the application scenario, a video surveillance system might be required to track targets over short sequences (e.g. in tightly-controlled and well-defined areas like entrances to buildings), or for very long sequences (e.g. in large and relatively constrained scenes like parking lots and shopping malls).

As indicated in the previous sections, multi-target tracking (e.g. Figure 3.1) is an extremely challenging problem. This is due to the uncontrolled nature of videos in which targets must be tracked and the necessity of coping with hard visual phenomena like uncontrolled and variable illumination that affects target appearance, target occlusion, and the need to track several targets in crowded scenes. Despite all these difficulties, a tracking system must be able to identify and track robustly and precisely different targets, and solve the data association problem. In each time instant the target identification implicates the detection, the appearance modeling, motion estimation, and location prediction. As above stated, a data association algorithm is the association of detected targets with one of the multiple tracked targets at each time.

Sparse methods that use regularized basis expansions to leverage a redundant dictionary of target templates for target discrimination have been recently applied to video surveillance problems like re-identification [55] and single-target tracking [3, 104]. These approaches are attractive in such systems because they effectively *defer* the decision about which target aspects are needed for discriminating new observations until the time of updating the tracker. Through regularization, they can leverage multiple aspects of targets by selecting the best ones to use when reconstructing a new observations. This approach has proven to be very robust for single-target tracking and person re-identification.

In previous chapter (Chapter 3), we defined a sparse approach to solve the data association problem using $\ell_1$-regularized basis expansions to associate detected targets to running trackers. However, this method has a number of limitations. All experiments, for example, assume that perfect detection information is available (i.e. that the locations, but not the identities, of targets is perfectly known). It exclusively focus on demonstrating the viability of solving the *data association problem* using sparse regularized basis expansions, and not on the solution of the entire end-to-end multi-target tracking problem.

In this chapter, we propose an end-to-end multi-target tracking system that uses the machinery of $\ell_1$-regularized basis expansions to solve both the target tracking and data association problems. Starting from noisy detections in video sequences, our approach attempts to reconstruct new observations using a regularized linear combination of tracklets (short target tracking sequences) already identified. An off-the-shelf pedestrian detector (the same used in [94]) is being used to acquire new detections. In particular, we define a novel global association phase (based on a weighted sparse reconstruction) that ensures long-term consistency of tracks.

The produced model is robust to occlusion and similarity appearance and allows to manage complex scenario. Data association is implemented via two different types of association phases: a local data association phase that considers at recent observations only, and a global association phase that looks at the entire sequence of associations up to a given time and enforces a long-term measure of association consistency. For this global data association, we exploit collected statistics of the very $\ell_1$-regularized basis expansions, that are used to perform local association.

In the next section (Section 4.1), we describe our local and global data association methodologies and present an energy formulation which is then optimized to solve the data association problem and define our end-to-end target tracking system. In Section 4.2 we report a number of experiments performed on publicly available multi-target tracking datasets. We conclude in Section 4.3 with a discussion of our contribution and some indication for future research directions.

## 4.1 Multi-Target Tracking by Sparse Reconstruction

This chapter extends the solution proposed in Chapter 3, that is based on $\ell_1$-regularization but uses ideal detections. Here a more realistic scenario using bounding boxes obtained from a pedestrian detector is being targeted. In the next section we formalize the multi-target tracking problem and introduce the notation used in the chapter, that for practical reasons, is equally to Chapter 3's formalization. After that, we introduce the local approach to solve the tracking-by-detection problem, defining how we represent each tracklet and how they are associated to new detections. We finally define a global data association phase that reduces fragmentation and maintains consistency for long-term tracking.

### 4.1.1 The Multi-Target Tracking Problem

The multi-target tracking problem is a complex computer vision problem, mostly because it consists of multiple sub-problems: detection, identity modelling, and data association. Each one of these problems is challenging in its own. Considering a video stream $\Psi$ which duration is $T \in \mathbb{N}$ frames, at each time $t$, $y_t$ denotes the set of $L_t$ different observations that are obtained from a pedestrian detector.

$Y$ evidences the complete set of observations over the whole video:

$$Y = \{y_t : t \in [1, T]\}, y_t = \{y_t^i\}_{i=1}^{L_t}. \tag{4.1}$$

Suppose there are $K$ different targets (persons) evolving in the scene during the whole video. A tracking algorithm has the aim of defining a set of tracklets:

$$\Omega = \{\omega_k : k \in [1, K]\}, \tag{4.2}$$

one for each target, avoiding both the creation of different tracklets for the same identity (i.e. fragmentation) and mixing two targets in a single tracklet (i.e. identity switch). Considering the complete set of observations $Y$, each tracklet $\omega_k$ will be characterized by a subset of observations. Each observation of $\omega_k$ belongs to a distinct time instant:

$$\omega_k = \{y_t^l : l \in [1, L_t], t \in [t_{ks}, t_{ke}]\} \subseteq Y, \tag{4.3}$$

where $t_{ks}$ is the time of first appearance for a target $k$ and $t_{ke}$ is the last appearance or *exit time* of target $k$ (hence $t_{ks} \leq t_{ke}$). Using this basic formulation of the tracking problem, we can now proceed to the definition of our data association algorithm that actually creates new tracklets and associates new observations with existing ones.

### 4.1.2 Local Data Association

In this section we will explain each step of our local data association model, giving an overview of the method in Algorithm 2. This component of our method is indicated as *local data association* to distinguish it from the *global data association* phase which will be later introduced. By "local" we mean that this data association phase only considers the current frame and current set of tracklets. Our local association relies on the solution of an $\ell_1$-regularized basis expansion problem to determine the possible association of new detections to existing tracklets. We maintain a discriminative basis for each tracklet, composed of the features computed from all detections associated to the tracklet $\omega_k$.

In order to be effective, the discriminative basis must be built using a powerful feature representation $\mathbf{f}(y)$ for each detection $y \in Y$. Once again, pedestrian detection is not the focus of this system, so the detection phase is hence provided by [94] on the datasets used to evaluate our algorithm. We use the powerful yet efficient feature representation proposed in [55], which extracts weighted histograms

**Algorithm 2:** Local data association

```
Data: $y_t$, $\mathbf{B}$, $\Omega$, $\tau$
/* Association error and spatial proximity */
foreach $y_t^i \in y_t$ do
    $\min_{\boldsymbol{\alpha}} \|\mathbf{f}(y_t^i) - \mathbf{B}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1$ ;
    compute $\varepsilon_k^i$, $s_k^i\,\forall k$ ;
    $a_k^i = \varepsilon_k^i + s_k^i\,\forall k$ ;
    if $a_k^i > \tau\ \forall k$ then
        $a_k^i = \infty$
    end if
end foreach
$\Omega_t = \Omega$ : local set of tracklets ;
/* Associate detections to tracklets */
while $\Omega_t \neq \emptyset \wedge y_t \neq \emptyset$ do
    $(\hat{k}, \hat{i}) = \arg\min_{k,i} a_k^i$;
    $\omega_{\hat{k}} = \omega_{\hat{k}} \cup \{y_t^{\hat{i}}\}$ ;
    $y_t = \{y_t \setminus y_t^{\hat{i}}\}$ ;
    $\Omega_t = \{\Omega_t \setminus \omega_{t,\hat{k}}\}$ ;
end while
/* New tracklets with remaining detections */
while $\bar{y}_t \neq \emptyset$ do
    $\Omega = \Omega \cup \bar{y}_t$
end while
update $\mathbf{B}$;
```

on overlapping stripes in the HS, RGB and LAB color domains as well as the Histogram of Oriented Gradients [20] (HOG) descriptor. Additional details about the feature representation are given in Section 4.2.

In the following subsections we will explain how we build a discriminative basis for local association and how we use a regularized sparse basis expansion to compute a reconstruction error according to the basis of each existing target. Finally the usage of this information to actually perform the data association will be presented.

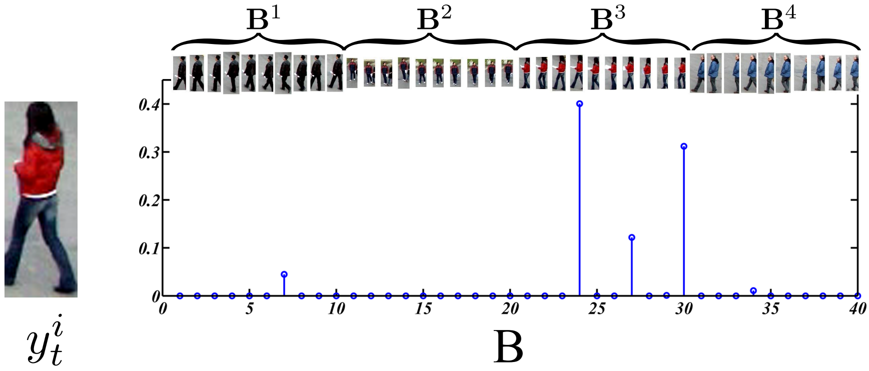


Figure 4.1: Sparse basis expansion. This figure illustrates how an observation is reconstructed in terms of a basis of tracklet observations. The $x$-axes indexes the coefficients $\boldsymbol{\alpha}$ in the reconstruction computed using Equation (4.7). Note how the reconstruction is *sparse* in that only a few basis vectors have non-zero $\boldsymbol{\alpha}$ coefficients.

### Discriminative Basis Construction

The discriminative basis $\mathbf{B}$ is an aggregation of all observed detections up to the current processing point, organized in a way that maintains the information about association between a detection and a tracklet. Specifically, it is composed of multiple sub-basis, one for each tracklet $\omega_k$:

$$\mathbf{B} = \Big[ \mathbf{B}^1, \ \ldots, \ \mathbf{B}^k, \ \ldots, \ \mathbf{B}^K \Big]. \tag{4.4}$$

At the first frame $t$ of a video $\Psi$ where detections are obtained, we use all the detections to define the initial set of sub-basis, one for each $y_t^i$ bounding boxes given by the detector. For example if there are $K = 3$ detections at the first frame $t = 1$, the discriminative basis $\mathbf{B}$ is being composed of 3 sub-basis, each one with only one template:

$$\mathbf{B} = \Big[ \mathbf{f}\left(y_1^1\right) \ \Big| \ \mathbf{f}\left(y_1^2\right) \ \Big| \ \mathbf{f}\left(y_1^3\right) \Big]. \tag{4.5}$$

After initialization, the algorithm continues to accumulate new detections by either associating them to existing tracklets, or creating a new tracklet if it is not possible to perform association. Hence at any time, each sub-basis $\mathbf{B}^k$ is composed of all the features extracted on all the detections associated to $\omega_k$ up to now:

$$\mathbf{B}^k = \left[ \mathbf{f}\left(y_{t_{ks}}^{i_{t_{ks}}}\right), \ldots, \mathbf{f}\left(y_t^{i_t}\right), \ldots, \mathbf{f}\left(y_{t_{ke}}^{i_{t_{ke}}}\right) \right] \ \forall\, y_t^{i_t} \in \omega_k. \tag{4.6}$$

### Regularized Sparse Basis Expansion

The discriminative basis $\mathbf{B}$ is used to perform a sparse reconstruction of an unknown detection. Our algorithm uses the $\ell_1$-regularization method to compute reconstruction coefficients. More precisely, the coefficient strength indicates which basis vectors of a tracked target contribute most to the sparse reconstruction. These coefficients are used by the method to verify how much affinity there is between the new detection $y_t^i$ and each tracklet $\omega_k$.

The upper part of Figure 4.1 depicts an example of a discriminative basis. The basis $\mathbf{B} \in \mathbb{R}^{m \times D}$ is called *discriminative*, because each vector of $\mathbf{B}$ is a feature vector $\mathbf{f}(y_t^{i_t}) \in \mathbb{R}^m$ associated with a specific target label, maintaining discriminative information about each sub-basis, where:

- $m$ is the dimension of each feature vector;
- $D_k = |\omega_k|$ are the number of detections associated with tracklet $\omega_k$;
- and $D = \sum_{k=1}^{K} D_k$ is the total number of templates in the discriminative basis.

We solve a $\ell_1$-regularized optimization problem:

$$\min_{\boldsymbol{\alpha}} \|\mathbf{f}(y_t^i) - \mathbf{B}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1, \tag{4.7}$$

to obtain the sparse reconstruction vector $\boldsymbol{\alpha} \in \mathbb{R}^D$. This vector is composed of $D$ coefficients that indicate how to best reconstruct the new observation $\mathbf{f}(y_t^i)$ using a linear combination of the sample vectors in $\mathbf{B}$. The parameter $\lambda \in \mathbb{R}^+$ in Equation (4.7) controls the sparsity of the reconstruction $\boldsymbol{\alpha}$: the larger the value of $\lambda$, the lower the $\ell_1$ norm of the reconstruction coefficients $\boldsymbol{\alpha}$.

One of these sparse reconstruction vectors $\boldsymbol{\alpha}$ is illustrated in Figure 4.1, where the magnitude of each reconstruction coefficients $\alpha_k^d$ is depicted for each basis vector. The coefficients $\alpha_k^d \neq 0$ indicate how much the corresponding template is used to reconstruct the detection. The larger the value of $\alpha_k^d$, the more the template is used to reconstruct the new observation. It must be noted that the coefficients of high magnitude are concentrated in the sub-basis $\mathbf{B^3}$ actually corresponding to the tracklet created for previous detections of the same person.

To estimate to which tracklet $\omega_k$ a new observation $y_t^i$ should be associated, we analyze the reconstruction coefficients in $\boldsymbol{\alpha}$. The vector $\boldsymbol{\alpha}$ can be seen as the concatenation of $\boldsymbol{\alpha}_k$ that are the coefficients corresponding to each sub-basis, such that

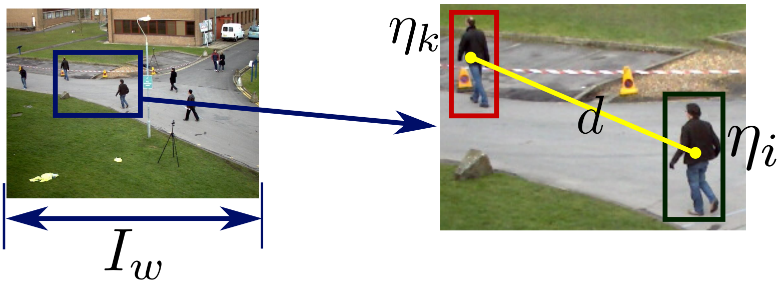


Figure 4.2: Example of the spatial proximity measure. This measure is obtained computing the distance $d$ between the two centroid observation and using the image's width $I_w$ to normalize $d$.

$\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1 \;\; \boldsymbol{\alpha}_2 \;\; \cdots \;\; \boldsymbol{\alpha}_k \;\; \cdots \;\; \boldsymbol{\alpha}_K]'$, where each $\boldsymbol{\alpha}_k$ is composed of $D_k$ coefficients. Note that $D_k$ can be different for each tracklet. To identify the associated target we define a *reconstruction error* $\varepsilon_k^i$ for each $(k, i)$:

$$\varepsilon_k^i = \|\mathbf{f}(y_t^i) - \mathbf{B}^k \boldsymbol{\alpha}_k\|_2. \tag{4.8}$$

The value $\varepsilon_k^i$ corresponds to the reconstruction error when $\mathbf{f}(y_t^i)$ is reconstructed using only those coefficients from $\boldsymbol{\alpha}$ and columns from $\mathbf{B}$ that correspond to the tracklet $\omega_k$. Since the feature is $\ell_2$ normalized, $\varepsilon_k^i \in [0, 1]$. The smaller the error $\varepsilon_k^i$, the greater the likelihood that $y_t^i$ is a new observation of the target corresponding to the tracklet $\omega_k$.

### Spatial Proximity

Since we are processing videos, a new detection of a target will most of the time appear near the last detection associated with its corresponding tracklet. To include spatial information during the tracking process, for each tracklet we compute the distance between the centroid of the most recent detection associated and the detections that should be associated (see Figure 4.2). We consider two vectors $\eta_k$ and $\eta_i$, where $\eta_k$ is composed of the position $(x_k, y_k)$ of the last detection associated to the tracklet $\omega_k$, and $\eta_i$ contains the coordinates of the centroid of a new observation.

The spatial proximity score between the tracklet $\omega_k$ and observation $y_t^i$ is:

$$s_k^i = \frac{\sqrt{(\eta_k - \eta_i)(\eta_k - \eta_i)'}}{I_w}, s_k^i \in [0, 1]. \tag{4.9}$$

This measure of spatial proximity is the euclidean distance between $\eta_k$ and $\eta_i$ normalized by the width of frame image ($I_w$). If the detections and the new observations are too distant, the score $s_k^i$ will tend to one, while its value will tend to zero if the distance is small.

To compute the association between tracklets and the new detections of frame $t$, we define the association error as a linear combination of the centroid difference and the reconstruction error $\varepsilon_k^i$ introduced in the previous section. Specifically, the association error $a_k^i$ is:

$$a_k^i = \varepsilon_k^i + s_k^i, \ \ \forall (k,i) \ \in \ [1,K] \times [1,L_t]. \tag{4.10}$$

**Putting it all together**

The association process, precisely defined in Algorithm 2, goes as follows. We first compute the association error, using Equation (4.10), for all detections of the current frame $t$. In the lines from 5 to 7 in the algorithm, $a_k^i$ is set to $\infty$ if its value is bigger than a threshold $\tau$ for all existing tracklets. This phase excludes cases in which a new detection $y_t^i$ is far (both in appearance and space) from any tracklet. The association between tracklets and detections is computed by a greedy approach (see lines 9 through 15). After the greedy association some detections may be unassociated, either due to the threshold $\tau$ or because they could not find a matching tracklet during the greedy association. In those cases, these leftover detections in $\bar{y}_t$ are used to define new tracklets.

It must be noted that we do not perform any pruning, but rather keep all detections in the discriminative basis. In this way we maintain a good up-to-date description of each target, but also all the appearance variations the target can have while evolving in the observed scene. In this way if we lose a target for some frames, it can be associate properly when it appears again even if the pose or appearance is similar only to old detections. In some cases the local association is not enough to re-acquire a target, and tracklets can be fragmented. To resolve these types of problems, in the next section we define an approach to enforcing long-term consistency in tracks.

### 4.1.3 Global Association for Long-term Consistency

During the multi-target tracking process multiple trajectories are created. It may therefore happen that more than one corresponds to the same subject. This problem is referred to as *tracklet fragmentation*, and to resolve these types of problems, state-of-the-art methods rely on some sort of global association that merges multiple tracklets of the same target. For example, in [3] the authors use a global association approach based on a measure of "tracklet confidence". This focuses on tracklets with low confidence, which are more likely to be fragmented. They are either associated with other tracklets with high confidence or with single detections not associated to any tracks, or they are terminated.

Our method to global association is based on the estimation of a compatibility score $C_k^q$ between existing $\omega_k$ and $\omega_q$ in a leave-one-out manner and merge them by a greedy pairing if this score is high. The computation of the compatibility scores is based on a weighted version of the $\ell_1$-regularized reconstruction already being used for local data association [12, 55]. We detail in Algorithm 3 our global association phase. It should be observed tracklets will merge only if their compatibility score exceeds a threshold $\gamma$ and exclude any tracklets joined in the previous step. This global association phase is performed at regular steps every $T_g$ frames, and iterated a fixed number of times (see the $iter$ parameter in Algorithm 3) to consider multiple merges of tracklets at one time.

### Weighted $\ell_1$-minimization

To estimate the compatibility between existing tracklets we prioritize the templates of each sub-basis that are used most frequently during the tracking process and can thus be considered as more representative. We compute a usage coefficient $A_k^d$ for each template of the discriminative basis $\mathbf{B}$, based on how often and with what magnitude it is used during the association process step. Each time a new detection is added to a tracklet $\omega_k$, the coefficient associated to the new detection is initialized as:

$$A_k^{D_k+1} = \frac{1}{D_k} \sum_{d=1}^{D_k} \frac{|A_k^d|}{\delta^d}, \tag{4.11}$$

where $A_k^d$ are the usage coefficients of each template $\mathbf{b}_d^k$ of $\mathbf{B}^k$ and $\delta^d$ indicates how many times a particular template $d$ of $\mathbf{B}$ has been used (i.e. that a detection was associated with the tracklet $k$ with a non null $\alpha_k^d$ for that template) during the entire tracking process up to the current time step. During the tracking process at each association step, each $A_k^d$ is incremented with the correspondence $\alpha_k^d$ value:

$$A_k^d = A_k^d + \left|\alpha_k^d\right|. \tag{4.12}$$

**Algorithm 3:** Global data association

```
Data: $\mathbf{B}$, $\Omega$, and $\gamma$
$\Omega_t = \emptyset$ : local set of tracklets;
for iteration = 1 : iter do
    /* Compatibility score matrix computation */
    foreach $k \in [1, K] \setminus \Omega_t$ do
        $\hat{\mathbf{B}} = \mathbf{B} \setminus \mathbf{B}^k$;
        compute $\hat{\mathbf{p}}$ for $\hat{\mathbf{B}}$;
        $c_q^k = 0,\ \forall q \in [1, K]$;
        foreach $d \in [1, D_k]$ do
            $\min_{\hat{\boldsymbol{\alpha}}} \|\mathbf{b}_d^k - \hat{\mathbf{B}}\hat{\boldsymbol{\alpha}}\|_2^2 + \lambda \|\mathrm{diag}(\hat{\mathbf{p}})\hat{\boldsymbol{\alpha}}\|_1$;
            $c_q^k \mathrel{+}= \sum_{j=1}^{D_q} \hat{\boldsymbol{\alpha}}_q^j,\ \forall q \in [1, K] \setminus \{k, \Omega_t\}$;
        end foreach
        $C_q^k = \frac{c_q^k}{\sum_{q=1, q \neq k}^{K} c_q^k},\ \forall q \in [1, K] \setminus \{k, \Omega_t\}$ ;
    end foreach
    /* Association between compatible tracklets */
    while $|\{C_k^q : C_k^q > \gamma\}| \neq \emptyset$ do
        $(\hat{q}, \hat{k}) = \arg\max_{k,q} C_k^q$;
        $\omega_{\hat{k}} = \omega_{\hat{k}} \cup \{\omega_{\hat{q}}\}$ ;
        $C_{\hat{k}}^{\hat{q}} = 0^+$ ; $C_{\hat{q}}^{\hat{k}} = 0^+$;
        $\Omega_t = \Omega_t \cup \{\hat{k}, \hat{q}\}$
    end while
end for
```

**Algorithm 4:** Multi-target tracking

```
Data: $\psi_t \in \Psi, T_g$
$\{y_t^i\}$ = detector($\psi_t$);
compute $\mathbf{f}(y_t^i)\,\forall y_t^i$ ;
/* Discriminative basis initialization */
if $\mathbf{B} = \emptyset$ then
    $\mathbf{B}$ = $[\mathbf{f}(\{y_t^i\}_{i=1}^{L})]$
end if
/* Global data association phase */
if $\mathrm{mod}(t, T_g) == 0$ then
    compute Algorithm 3;
end if
/* Local data association phase */
compute Algorithm 2;
```

Then $u_k^d$ is computed as:

$$u_k^d = \frac{A_k^d}{\delta^d}, \forall d \in D_k. \tag{4.13}$$

We normalized the $u_k^d$ value with a min-max normalization in order to scale each coefficient of the same sub-basis, and finally mean-center it to obtain $\overline{u}_k^d$. The idea of weighted $\ell_1$-regularization is to use weights to prioritize (or penalize) some templates over others in the reconstruction. In our case, the weight should be inversely proportional to $\overline{u}_k^d$. We hence define $p_k^d$ as a sigmoid:

$$p_k^d = \frac{1}{1 + e^{-6 \cdot \overline{u}_k^d}}. \tag{4.14}$$

In the global data association Algorithm 3, considering a tracklet $\omega_k$ for which we want to estimate its compatibility scores with the other tracklets, we define the restricted discriminative basis $\hat{\mathbf{B}} = \mathbf{B} \setminus \mathbf{B}^k$. After that, the vector $\hat{\mathbf{p}}$ is built by concatenating all weights of the templates present in $\hat{\mathbf{B}}$. For each template $d$ in $\omega_k$, we compute the following weighted $\ell_1$-minimization:

$$\min_{\hat{\boldsymbol{\alpha}}} \|\mathbf{b}_d^k - \hat{\mathbf{B}}\hat{\boldsymbol{\alpha}}\|_2^2 + \lambda \|\mathrm{diag}(\hat{\mathbf{p}})\hat{\boldsymbol{\alpha}}\|_1. \tag{4.15}$$

Each compatibility score $C_k^q$ between tracklets $\omega_k$ and $\omega_q$, is defined as the sum of the coefficients $\hat{\alpha}_q^d$ that correspond to the sub-basis $\hat{\mathbf{B}}^q$. This coefficient is further

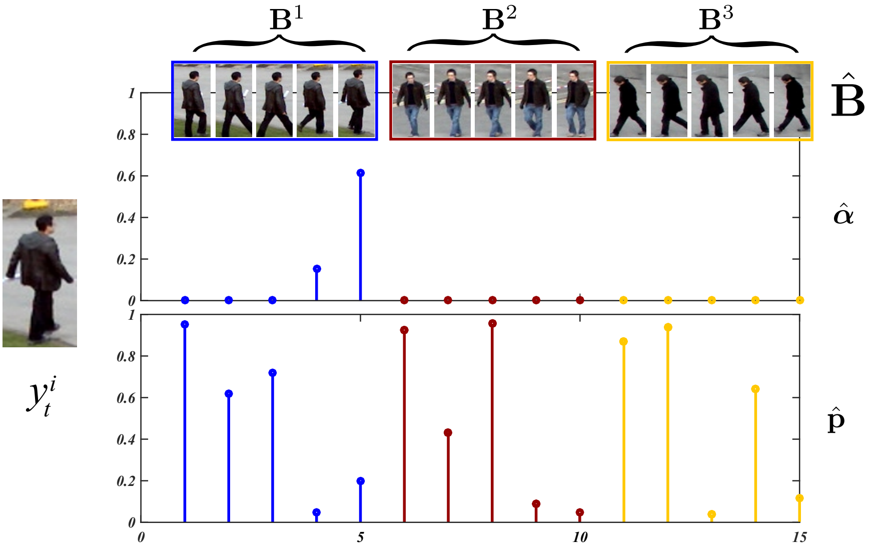
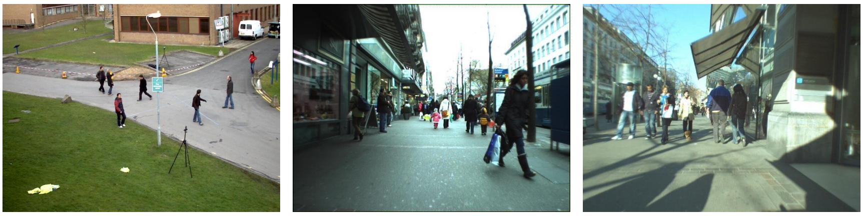


Figure 4.3: Example of the $\hat{\alpha}$ vector obtained with the use of weighted sparse reconstruction (Eq. (4.15)). In this case we consider a discriminative base $\hat{\mathbf{B}}$ composed by three targets and a new observation $y_t^i$. The vector $\hat{\mathbf{p}}$ is obtained using the Equation (4.14) of the proposed method. Each coefficient $p_k^d$ is obtained considering the usage of each vector template $\mathbf{b}_d^k \in \hat{\mathbf{B}}$.

normalized by dividing by the sum over all $q$ being compared, see lines 4-11 in Algorithm 3 for the complete process of computing compatibility scores with the weighted $\ell_1$-minimization.

In Figure 4.3 we illustrate an example of weighted sparse reconstruction. We consider a case in which after five frames, $\hat{\mathbf{B}}$ consists of three sub-basis, where each of them has five templates $\hat{\mathbf{b}}_d^k$. Considering the local association process until the sixth frame, we obtain the $\hat{\mathbf{p}}$ vector shown in figure. Each coefficient $p_k^d$ has a different value which depends on how each template has been used in the previous step. The $\hat{\alpha}$ vector shown is obtained reconstructing the observation $y_t^i$ presented in figure with the Equation (4.15). How we can see for the sub-basis $\mathbf{B}^1$, a high value of $p_k^d$ correspond a very low value of $\alpha_k^d$. On the other hand, with a low value of $p_k^d$, we have a high value of $\alpha_k^d$. The $\alpha_k^d$ associated to the other target are equally to zero, due to the regularization of the method, while the $p_k^d$ value are not zero because each template is used differently in each previous step.



(a) *s2.l1-view01* (PETS dataset). (b) *Bahnhof* (ETH Dataset). (c) *Sunnyday* (ETH dataset).

Figure 4.4: Sample frames from the PETS and ETH datasets.

The local and global data association phases are combined in our end-to-end tracking-by-detection given in detail in Algorithm 4. In the next section we analyze the results obtained with the proposed method on two different datasets.

## 4.2 Experimental Results

In this section we report on a set of experiments performed using our proposed method and a comparison with two state-of-the-art methods [3, 94]. Experiments were conducted using the *s2.l1-view01* sequence of the PETS 2009 dataset [1] and the: *Bahnhof* and *Sunnyday* sequences from the ETH Dataset [25]. Some sample frames are shown in Fig. 4.4. For a fair comparison, as input of our method we use the detections also used in [94].

For the target appearance representation, we use the descriptor proposed in [55, 56]. In particular, each detection $y_t^i$ is scaled to a canonical resolution of $128 \times 64$ pixels and build a spatial pyramid by dividing the image into overlapping horizontal stripes of 16 pixels in height. From each stripe Hue-Saturation(HS), RGB and Lab histograms are extracted; background information is filtered using an Epanechnikov kernel. To these color histograms a HOG descriptor, computed excluding a border of 8 pixel from each side of the image, is concatenated. This descriptor has been shown to be robust to pose and illumination variations and has provided top state-of-the-art performance for re-identification in [50, 55, 56].

In the following, we first describe the datasets used for our experiments and then the results we obtained.

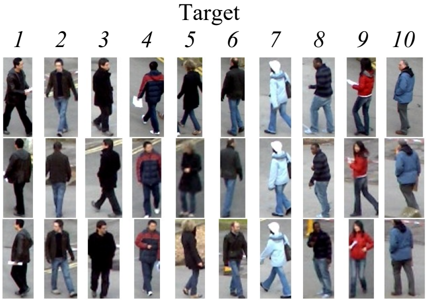


Figure 4.5: Sample detections for the *s2.l1-view01* sequence of the PETS and dataset.

### 4.2.1 Datasets and Metric

In Table 4.1 we report the main characteristics of each dataset. The *s2.l1-view01* sequence of the PETS dataset consists of 794 frames, acquired with a frame rate of 7fps. There are 10 targets in total, all having a different color appearance and different types of behavior. Moreover, some targets enter and leave the scene several times. Sample detections are shown in Fig. 4.5.

The ETH sequences, *Sunnyday* and *Bahnhof*, consist of 354 and 1000 frames respectively. Each sequence was acquired with a frame rate of 15fps. The difficulties in these two sequences are related to the crowded nature of the scenes. There are 124 targets in total that mutually occlude each other several times.

Performance of the proposed solution and state-of-the-art methods are reported using the USC metric [94], that is defined by:

- **Precision:** evaluation of the correctness of each tracklet;
- **Recall:** evaluation of the completeness of the created tracklets, considering the available detections;
- **Mostly tracked trajectories (MT):** measurement of the ratio of trajectories that are successfully tracked for more than 80% of their frames;
- **Partially tracked trajectories (PT):** measurement of the ratio of trajectories which are successfully tracked between 20% and 80%;
- **Mostly lost trajectories (ML):** measurement of the ratio of trajectories which are successfully tracked for less than 20%;
- **False positive per frame (FPF):** measurement of the average number of false positive detections per frame;

| | **PETS** | **ETH** [25] | |
|---|---|---|---|
| | *s2.l1-view01* | *Sunnyday* | *Bahnhof* |
| **# Persons** | 10 | 30 | 94 |
| **# Frames** | 794 | 354 | 1000 |
| **Frame Rate** | 7 fps | 14/15 fps | 14/15 fps |
| **Resolution** | $768 \times 576$ | $640 \times 480$ | $640 \times 480$ |
| **Difficulty** | Medium | Medium-Hard | Hard |

Table 4.1: Characteristics of the PETS and tracking sequences.

- **ID switch (IDS):** indication of the number of times that a tracked tracklet changes its id;
- **Tracklet fragmentation (Frags):** indication of the number of times that a groundtruth tracklet is interrupted;

As additional information we report about the number of tracks present in the groundtruth (GT) and the number of tracklets created by the approach (SO).

In the following we report on the performance of the proposed approach with and without global association and in comparison with state-of-the-art methods. In all our experiments we set $\gamma = 0.7$ and $T_g = 150$ for the global association phase.

### 4.2.2 Performance on PETS

In Table 4.2 we report the results obtained by the algorithm on the *s2.l1-view01* sequence of the PETS dataset. Specifically, we report results obtained with local association and with local+global association which are illustrated in comparison with state-of-the-art methods. From the table we see that the global association phase helps to merge fragmented tracks produced during the local association, and this results in a higher value of mostly tracked targets and a lower number of ID switches and fragmentations.

With respect to the state-of-the-art, our algorithm with local+global association outperforms the techniques of [3, 94]. However, it does result in a higher number of track fragmentations. In fact a fragmentation is generated every time a track is interrupted. We have observed that in our results some tracks are merged after the global association. In fact, some of these merged tracks contain gaps because of missing or unassociated detections. Moreover, the solutions in [3, 94] consider a ground truth composed by 23 and 19 tracks, respectively, because they consider

| Method | Recall | Precision | MT | PT | ML | FPF | IDS | Frags | GT | SO |
|---|---|---|---|---|---|---|---|---|---|---|
| Our (Local) | 88.30% | **97.44%** | 70.00% | 30.00% | 0.00% | **0.13** | **0** | 70 | 10 | 74 |
| Our (Local+Global) | **96.59%** | 94.12% | **100.00%** | **0.00%** | **0.00%** | 0.35 | 1 | 24 | 10 | 18 |
| Yang et al. [94] | 93.00% | 95.30% | 89.50% | 10.50% | 0.00% | 0.27 | **0** | 13 | 19 | - |
| Bae et al. [3] | - | - | **100.00%** | **0.00%** | **0.00%** | - | 4 | **4** | 23 | - |

Table 4.2: Performance on the *s2.l1-view01* of the PETS dataset.

| Method | Recall | Precision | MT | PT | ML | FPF | IDS | Frags | GT | SO |
|---|---|---|---|---|---|---|---|---|---|---|
| Our (Local) | 69.53% | 89.84% | 50.68% | 41.67% | 7.66% | **0.52** | 20 | 79 | 124 | 143 |
| Our (Local+Global) | 74.46% | 85.19% | 61.74% | 27.94% | 10.32% | 0.85 | **11** | 61 | 124 | 178 |
| Yang et al. [94] | **79.00%** | **90.4%** | 68.00% | 24.8% | 7.20% | 0.64 | **11** | **19** | 125 | - |
| Bae et al. [3] | - | - | **73.81%** | **23.81%** | **2.38%** | - | 18 | 38 | 126 | - |

Table 4.3: Performace on the *Sunnyday* and *Bahnhof* sequences of the ETH dataset.

those tracks that leave and re-enter the scene as different identities. Our experiments are instead evaluated considering the real identities.

The global association phase helps to merge fragmented tracks produced during local association, and this results in a higher value of mostly tracked targets and a lower number of ID switches and fragmentations with respect to the usage of local association only.

### 4.2.3 Performance on ETH

In Table 4.3 we report the results obtained by the proposed approach on the *Sunnyday* and *Bahnhof* sequences of the ETH dataset. First we report the results obtained by our method with local association and with local+global association. Then we show the comparison with the state-of-the-art [3, 94]. Results on this dataset are usually reported as the average of the two sequences.

With respect to the state-of-the-art, on one hand our method does yield fewer identity switches. On the other hand, our technique also yields generally worse results. This is mainly due to the *Bahnhof* sequence that has a larger number of persons to track and critical illumination conditions. In particular, in this sequence, as also for the PETS, tracks merged after the global association can have gaps due to missing or unassociated detections and this results in a higher number of fragmentations and a lower value of mostly tracked (MT).

## 4.3 Discussion and Conclusions

In this chapter we propose an end-to-end multi-target tracking approach based on $\ell_1$-regularized basis expansions to solve both the target tracking and data association problems. The use of sparse reconstruction allows the tracking approach to integrate observations of target appearance from the entire lifetime of each target. This renders the local data association phase robust to occlusion and changing target appearance. Moreover, we introduce a novel global data association phase that exploits the accumulated statistics of sparse local associations in order to merge tracklets likely to correspond to the same target.

The algorithm outperforms the state-of-the-art on the PETS sequence while maintaining comparable performance on the ETH dataset. In both cases the proposed solution results in few identity switches. We feel that sparse basis expansions are a powerful tool for robust appearance modeling and data association. Our ongoing research is focused on integrating global and local data association into a single, continuous framework that obviates the need to run global association at arbitrary intervals during the tracking process. Moreover, as observed from the experimental results it is worth investigating a solution to the gaps created by global association by looking at those detections that remain unassociated after local and global data association.

With the aim to estimate facial characteristic which include Gender, Age, Ethnicity and head pose, in the next chapter a method to joint estimate these ones is defined using a random forest base algorithm.

# Chapter 5

# Multi-Objective Random Forests for Multiple Facial Characteristic Estimation

*The difference between the impossible and the possible*
*lies in a person's determination.*

- Thomas Charles "Tommy" Lasorda

Facial characteristics like Gender, Age, Ethnicity and face pose, are important to estimate in many computer vision applications. Estimating the Gender and Age can be used to adapt advertising displayed on nearby screens, and pose estimation can allow users to interact with devices by simply looking at them. Although head pose is not a biometric characteristic, it is related to the gaze of a person and therefore can be an important aspect of behaviour and social interaction understanding. Despite the attention received in recent years, estimation of multiple face characteristics, and especially multiple soft biometrics like Age, Gender and Ethnicity, remains a difficult problem and an active area of research in the computer vision community.

As extensively discussed in Section 2.2, most systems for estimating characteristics like Age, head pose [34,44] and Gender [54,103] use their own sets of custom

Part of this chapter has been published as "MORF: Multi-Objective Random Forests for Face Characteristic Estimation" in Proceedings of *International Conference on Advanced Video- and Signal-based Surveillance (AVSS)*, August, Karlsruhe, Germany, 2015 [22].

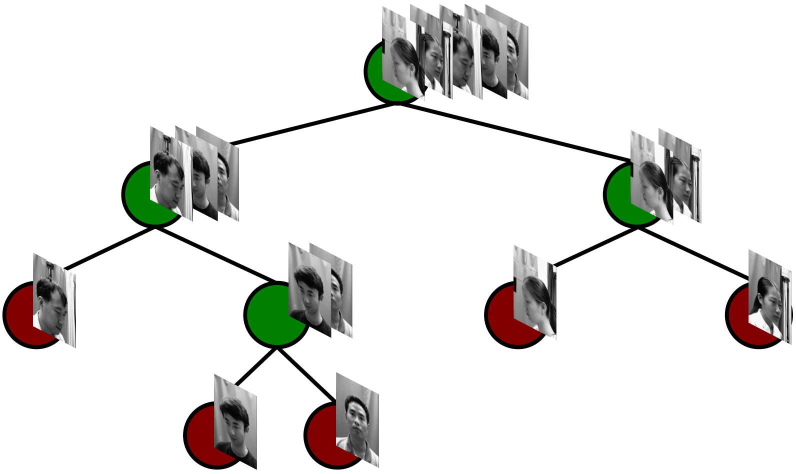


Figure 5.1: A multi-objective random tree. In this example, the tree tends to split training images on Gender at the higher levels and on pose angles at the lower ones.

features and specific estimation techniques. In systems requiring simultaneous estimates of target characteristics in real-time, this can be wasteful since much work (like feature extraction) is duplicated. Moreover, it may be easier to estimate one characteristic after conditioning it on a subset of others. Instead of estimating characteristics individually, we believe that a single system can be built in order to be able to perform a joint estimation of all these characteristics using a single pool of features and a single estimator. In this way, estimation of multiple characteristics can be made more efficient and more robust.

We believe that the usage of random decision forests [18] can contribute to provide a unified framework for multi-objective estimation. In this chapter we show how they can be used to simultaneously estimate multiple characteristics using a single pool of features. We propose a new information gain formulation enabling the use of multiple (potentially heterogeneous) characteristics to train a random forest. We demonstrate the effectiveness of the proposed approach for jointly estimating head pose, Gender, Age and Ethnicity from single face images. As far as head pose is concerned, Pan angles have been mainly taken into account and a more accurate review is proposed in the following sections. Fig. 5.1 illustrates the main idea behind Multi-Objective Random Forests (MORF). Early levels of each random tree will tend to specialize on a subset of characteristics, effectively conditioning later levels on them.

In this chapter we describe a technique for joint estimation of head pose and multiple soft biometrics from faces (Age, Gender and Ethnicity). Our proposed Multi-Objective Random Forests (MORF) framework is a unified model for the joint estimation of multiple characteristics that automatically adapts the measure of information gain used for evaluating the quality of weak learners. Since facial characteristics are related in the feature space, estimating all of them jointly can be beneficial as trees can learn to condition the estimation of some characteristics on others. We reformulate the splitting criterion of random trees in the our multi-objective formulation and evaluate it on publicly available face characteristic estimation imagery. These preliminary experiments show promising results.

In the next section we discuss the relevant works on soft biometric and head pose estimation from face images based on the random forests. In Section 5.2 we describe our approach to multi-objective estimation with random decision forests. We report on a series of experiments performed to evaluate the potential of our proposed estimation framework in Section 5.3 and we conclude in Section 5.4 with a discussion of our method contribution.

## 5.1 Random Forests

*Random Decision Forests*, or simply random forests, are a learning method used for different aims, as classification or regression. A random forest is composed by a set of decision random trees (potentially many), where each of them gives a different evaluation in the estimation process. The training process is built in the same way in each type of application, processing data differently, depending on their distribution (discrete or continuous).

The training phase is performed with the same procedure in every forest tree, but independently from each other. Considering a classification case, assume a set $\mathcal{S}_0$ of $N$ train elements:

$$\mathcal{S}_0 = \{(\mathbf{x}_i, y_i) \mid i = 1, \ldots, N\}, \tag{5.1}$$

where:

- $\mathbf{x}_i = (x_1, x_2, \ldots x_d) \in \mathbb{R}^d$ are the data with which train a forest;
- $y_i \in \mathcal{Y}$ is the set of labels of each train data.

Each random tree is built by recursively splitting this initial set $\mathcal{S}_0$ of labeled examples in such a way that the mutual information between the set of examples at each node and the label $y_i$ being estimated is maximized.

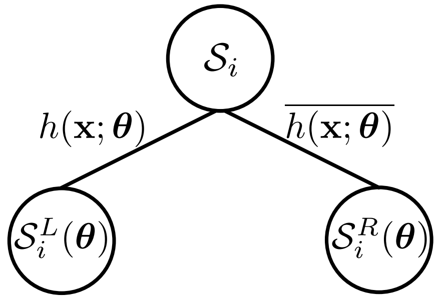

$\mathcal{S}_i$
$h(\mathbf{x};\boldsymbol{\theta})$
$\overline{h(\mathbf{x};\boldsymbol{\theta})}$
$\mathcal{S}_i^L(\boldsymbol{\theta})$
$\mathcal{S}_i^R(\boldsymbol{\theta})$

Figure 5.2: The image illustrates an exemplification of the split process for a generic node of a tree in a random forest.

In each node of each tree a same splitting function $h(\mathbf{x};\boldsymbol{\theta})$ is defined that splits the set in two child sets. The split function plays a crucial role both in training and testing phase, because it determines the quality of the classification. It is usually performed with a simple geometric parametrization, but it can consider a specific model, depending on the data distribution. The more simple parametrization is the linear one that can be defined as:

$$h(\mathbf{x};\theta_1,\theta_2) = \begin{cases} 1 & \text{if } x_{\theta_1} > x_{\theta_2} \\ 0 & \text{otherwise.} \end{cases} \tag{5.2}$$

Given a parametrization $\boldsymbol{\theta} = (\theta_1, \theta_2)$ of the splitting function, the left and right child sets are defined (Fig. 5.2) of $\mathcal{S}_i$ as:

$$\begin{aligned} \mathcal{S}_i^L(\boldsymbol{\theta}) &= \{\mathbf{x} \in \mathcal{S}_i \mid h(\mathbf{x};\theta_1,\theta_2) = 1\} \\ \mathcal{S}_i^R(\boldsymbol{\theta}) &= \{\mathbf{x} \in \mathcal{S}_i \mid h(\mathbf{x};\theta_1,\theta_2) = 0\}\,. \end{aligned} \tag{5.3}$$

By construction, the splitting function $h(\mathbf{x};\boldsymbol{\theta})$ guarantees that $\mathcal{S}_i^L(\boldsymbol{\theta})$ and $\mathcal{S}_i^R(\boldsymbol{\theta})$ partition $\mathcal{S}_i$. Trees are built by randomly sampling the parameter space $\boldsymbol{\theta} = (\theta_1, \theta_2)$ at each node $\mathcal{S}_i$ in each tree. At each internal node, $T$ random parameters $\boldsymbol{\theta}_t$ for $t = \{1, \ldots, T\}$ are generated, choosing the one $\boldsymbol{\theta}_t^*$ that produce the best split.

With respect to the available literature, we note that in order to evaluate the best parametrization, a measure that consider the entropy distribution in the resulting subsets with respect to the label being estimated is often used. Considering an internal node characterized by a subset $\mathcal{S}_i$, the common measure used is the *information gain*, defined as:

$$I_i(\boldsymbol{\theta}) = H(\mathcal{S}_i) - \sum_{j \in \{L,R\}} \frac{|\mathcal{S}_i^j(\boldsymbol{\theta})|}{|\mathcal{S}_i|} H(\mathcal{S}_i^j(\boldsymbol{\theta})), \tag{5.4}$$

where $H(\mathcal{S})$ is the label entropy in set $\mathcal{S}$:

$$H(\mathcal{S}) = -\sum_{y \in \mathcal{Y}} p(y|\mathcal{S}) \log p(y|\mathcal{S}), \tag{5.5}$$

for the discrete estimate of $p(y|\mathcal{S})$ from set $\mathcal{S}$:

$$p(y|\mathcal{S}) = \frac{|\{\mathbf{x}_i \in \mathcal{S} \mid y_i = y\}|}{|\mathcal{S}|}. \tag{5.6}$$

In this case the resulting split is chosen considering the highest information gain:

$$\boldsymbol{\theta}_i^* = \arg\max_t I_i(\boldsymbol{\theta}_t). \tag{5.7}$$

The usage of information gain to evaluate splits produces trees in which the entropy of the label distributions associated with the nodes decreases when descending in the tree, and thus prediction confidence increases. The number of $T$ tests controls the randomness of the resulting tree. Clearly, if $T = d(d-1)$ the best overall split at each node will be selected and there will be no randomness in the resulting tree.

During the forest creation, child nodes are added until a maximum depth $D$ is reached in each tree, or until a minimum number of training elements remains in the set $\mathcal{S}_i$ at node $i$. Independently from the purpose, the training phase is performed with the same structure, which is used to classify an unlabeled element $\mathbf{x}$. This can be done using the leaf label distribution to estimate a possible label, for instance, or by the average value between the distribution or by the max value between all results. Compared with the classification purpose, the regression case considers a continuous distribution, and a different formulation of the information gain. For a complete review on the random decision forest configuration, see the work presented in [18].

Thanks to its hierarchical structure, the random forests appear to be an efficient and fast way to define good computer vision application, splitting complex problems into a hierarchy of simpler ones. The idea is that each decision tree can give different results, but the forest resulting by combining all tree outputs can produce higher accuracy. According to the state-of-the-art, existing approaches based on random forests focus on a single characteristic estimation, or consider multiple characteristics independently. In the following section we shortly introduce these approaches with some relevant comments.

### 5.1.1 Random Forests for Face Characteristic Estimation

The random forest model has been extensively used for head pose estimation as well [18, 44]. In particular, in [44] the authors use a random forest for head pose estimation, implementing a version of Linear discriminant analysis (LDA) as a split function and evaluating their approach to estimate the pan or tilt angle separately. Fanelli *et al.* [27] propose an algorithm for estimating the head pose in RGB-D images considering a 3D model of the face, the 3D coordinates of the nose tip and the angles of rotation of a range image of a head. They use a random regression forest to estimate the pose, using the image of faces annotated with the nose-tip position. Their experiments show that the approach can handle real data presenting large pose changes, partial occlusions, and different facial expressions.

The random forests are also used for feature selection purpose as in [15, 40, 97] or in data mining [14, 78]. In [40] a variable section approach based on the random forests is proposed, that can distinguish between relevant and irrelevant variable in a general classification approach. Yaqub *et al.* [97] propose a technique that use random forests for 3D segmentation objects. They use the random forests to reduce the number of features in the classification approach, maintaining the relevant ones. In this way the object descriptors used in the classification approach is reduced improving the performance of the method.

The authors in [15] improve the Hough Forest [30] approach, based on random forests, defining a faster computation of the features and a faster evaluation of the learned model with minimal loss in accuracy. They define an algorithm for object detection on wearable device, that consider a part model detection approach. In this method the random forests are defined to reduce the number of parts for object detection allowing a real-time detection.

Due to his hierarchical structure, random forests are used in those cases where the train phase considers sets with many elements. In [14] the authors review the relevant work that use random forests to analyse genomic data, considering for example classification data, or to predict the gene-gene interaction. With the same consideration to analyse big data, Ristin *et al.* [78] define an approach for image classification considering all those dataset, like ImageNet [21], that are composed by millions of images and several classes. The authors proposed two different configuration of the random forests, considering two different split functions, one based on the distance between the class, and the second one based on a binary version of the SVM approach.

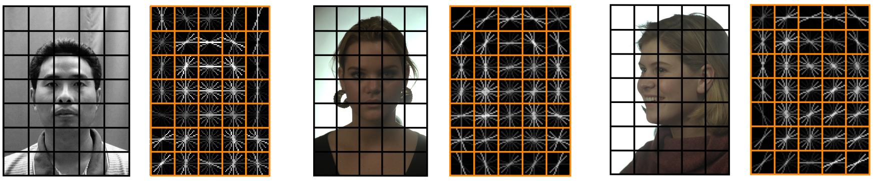


Figure 5.3: Three different face images and their corresponding HOG descriptors are depicted.

## 5.2 Multi-objective Random Forests (MORF)

In this section we describe our multi-objective estimation technique using random forests. First of all we introduce the feature description used for multi-objective face characteristic estimation.

### 5.2.1 Feature Representation for Faces

We use the HOG feature descriptor [20] to describe faces. It is a robust feature that describes orientation information and maintains some spatial information. All face images are first resized to a canonical dimension of $41 \times 54$ pixels, maintaining the aspect ratio of the images. Using the standard cell configuration of $8 \times 8$ pixels, the gradient orientations are extracted in each cell and quantized into 8 orientations. Block normalization is then applied to $4 \times 4$ blocks of cells, and normalized outputs are concatenated to the final HOG vector of 1024 bins. This feature is being used in a random forest framework to estimate four facial characteristics: head pose angle, Gender, Age and Ethnicity. See Figure 5.3 for an illustration of our HOG configuration.

### 5.2.2 Random Forests for Supervised Estimation

In this section we describe the basic single-objective estimation model upon which MORF are based and delineate how we extend it to multi-objective estimation model using random forests. In the following sections we use the same notation introduced in Section 5.1.

### Training Single-objective Random Trees

We consider supervised estimation problems and assume that a set of $N$ face images is available, where each of them is represented as a $d$-dimensional vector and labeled with a single characteristic $c$:

$$\mathcal{S}_0 = \{(\mathbf{x}_i, y_i) \mid i = 1, \ldots, N\}, \tag{5.8}$$

such that $\mathbf{x}_i = (x_1, x_2, \ldots x_d)$ and $y_i \in \mathcal{Y}^c$, the set of labels of characteristic $c$. A random tree is built by recursively splitting this initial set $\mathcal{S}_0$ of labeled examples in such a way that the mutual information between the set of examples at each node and the characteristic $c$ being estimated is maximized.

As previously mentioned, we define a splitting function that compares two dimensions in the face descriptor:

$$h(\mathbf{x}; \theta_1, \theta_2) = \begin{cases} 1 & \text{if } x_{\theta_1} > x_{\theta_2} \\ 0 & \text{otherwise.} \end{cases} \tag{5.9}$$

Given a parametrization $\boldsymbol{\theta} = (\theta_1, \theta_2)$ of the splitting function, we define the left $\mathcal{S}_i^L(\boldsymbol{\theta})$ and right $\mathcal{S}_i^R(\boldsymbol{\theta})$ child sets of $\mathcal{S}_i$, as well as reported in Eq. (5.4). By construction, the splitting function $h(\mathbf{x})$ guarantees that $\mathcal{S}_i^L(\boldsymbol{\theta})$ and $\mathcal{S}_i^R(\boldsymbol{\theta})$ partition $\mathcal{S}_i$. The quality of splits is measured by the *information gain* in the resulting subsets with respect to the characteristic $c$ being estimated:

$$I^c(\mathcal{S}_i, \boldsymbol{\theta}) = H_c(\mathcal{S}_i) - \sum_{j \in \{L,R\}} \frac{|\mathcal{S}_i^j(\boldsymbol{\theta})|}{|\mathcal{S}_i|} H_c(\mathcal{S}_i^j(\boldsymbol{\theta})), \tag{5.10}$$

where $H_c(\mathcal{S})$ is the entropy in set $\mathcal{S}$, considering the characteristic $c$:

$$H_c(\mathcal{S}) = - \sum_{y \in \mathcal{Y}^c} p(y|\mathcal{S}) \log p(y|\mathcal{S}). \tag{5.11}$$

As indicated in Section 5.1, using information gain to evaluate splits produces trees in which the entropy of the class distributions associated with the nodes decreases when descending in the tree, and thus prediction confidence increases.

Trees are built by randomly sampling the parameter space $\boldsymbol{\theta} = (\theta_1, \theta_2)$ at each node $\mathcal{S}_i$ in the tree. At each internal node, $T$ random parameters $\boldsymbol{\theta}_t$ for $t = \{1, \ldots, T\}$ are generated, and the split resulting in the highest information gain is chosen:

$$\boldsymbol{\theta}_i^* = \arg\max_t I_i(\boldsymbol{\theta}_t). \tag{5.12}$$

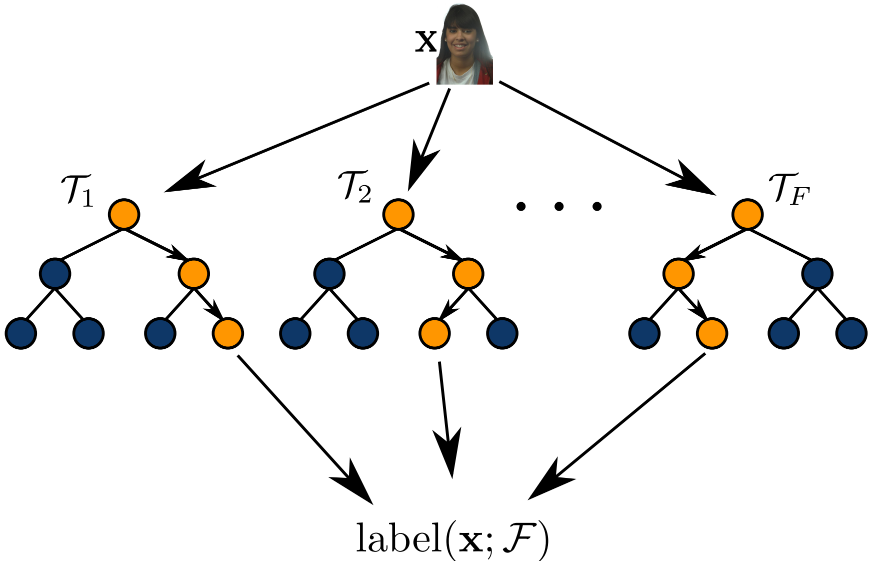


Figure 5.4: Random forests' estimation process. Considering an unlabelled image $\mathbf{x}$, in each tree $\mathcal{T}_i$, $\mathbf{x}$ follows a certain path (orange path) until leaf nodes. Each leaf distribution is used to estimate the label $\text{label}(\mathbf{x}; \mathcal{F})$

where $T$ controls the randomness of the resulting tree. Child nodes are added until a maximum depth $D$ is reached in the trees, or until a minimum number of training elements remains in the set $\mathcal{S}_i$ at node $i$.

### Estimation with Forests of Random Trees

A random tree $\mathcal{T}$ is thus defined by the split parameters at each internal node:

$$\mathcal{T} = \{\boldsymbol{\theta}_i\}_{i=1}^{i=|\mathcal{T}|}. \tag{5.13}$$

To ensure diversity and avoid overfitting possible with a single tree, a forest of trees is defined as $\mathcal{F} = \{\mathcal{T}_1, \ldots, \mathcal{T}_F\}$. The trees in $\mathcal{F}$ are trained independently and their outputs averaged to provide some sort of regularization.

Given an unlabeled test sample $\mathbf{x}$, we use the hierarchy of tests defined by each $\mathcal{T}$ to determine to which leaf node it arrives (in Figure 5.4 is illustrated an exemplification of the estimation process). Denoting the leaf node that $\mathbf{x}$ arrives to in tree $\mathcal{T}$ as $l(\mathbf{x}; \mathcal{T})$, we can estimate the unknown label $y$ of $\mathbf{x}$ using the training elements $\mathbf{x}_i$ that arrive to the same leaf node:

$$\text{label}(\mathbf{x};\mathcal{F}) = \arg\max_y \frac{1}{|\mathcal{F}|}\sum_{\mathcal{T}\in\mathcal{F}} p(y|L(\mathbf{x};\mathcal{T})), \tag{5.14}$$

where $L(\mathbf{x};\mathcal{T})$ is the set of training examples in the same leaf as $\mathbf{x}$ in tree $\mathcal{T}$:

$$L(\mathbf{x};\mathcal{T}) = \left\{(\mathbf{x}_i, y_i) \in \mathcal{S}_0 \mid l(\mathbf{x}_i;\mathcal{T}) = l(\mathbf{x};\mathcal{T})\right\}. \tag{5.15}$$

### Multi-objective Estimation with Random Forests

We now assume that each training sample $\mathbf{x}_i$ is labeled with $C > 1$ characteristics we wish to estimate:

$$\mathcal{S}_0 = \{(\mathbf{x}_i, \{y_i^j\}_{j=1}^C) \mid i = 1\ldots N\}. \tag{5.16}$$

Fig. 5.1 illustrates the main idea of our algorithm using the example of simultaneously estimating Gender and head pose. If constructed properly, the tree model should learn how to condition the estimation of one characteristic on the estimation of another, thus simplifying the problem. The fundamental difference between single-objective and multi-objective estimation is in how the information gain driving the splitting process is defined. In particular, there is no guarantee that the information gain in one characteristic is comparable in scale with the information gain in another.

We define a new normalized measure of information gain for multi-objective random forests, it weights the information of each characteristic $c$ by the ratio between the local entropy $H_c(\mathcal{S}_i)$ in $\mathcal{S}_i$ with respect to the root entropy $H_c(\mathcal{S}_0)$ in $\mathcal{S}_0$. This *locally weighted information gain* $I_{\text{lw}}$ is defined as:

$$I_{\text{lw}}(\mathcal{S}_i, \boldsymbol{\theta}) = \max_c \frac{H_c(\mathcal{S}_i)}{H_c(\mathcal{S}_0)} I^c(\mathcal{S}_i, \boldsymbol{\theta}). \tag{5.17}$$

The main idea behind the definition of $I_{\text{lw}}$ is to update weights, i.e. the ratio between the two entropy values, during the training process in order to scale each characteristic information gain based on how much entropy remains at the current depth.

During training, our method selects the split function parametrization $\boldsymbol{\theta}$ which maximizes the *locally weighted information gain* $I_{\text{lw}}$ according to one characteristic. According to the proposed information gain function, in each node of each tree the characteristic $c$ is chosen automatically and independently from the other ones. In this case, the estimation of each characteristic label $y^c$ of an unlabelled

| | **FERET** [72] | **CAS-PEAL-R1** [32] | **MIX Dataset** |
|---|---|---|---|
| **# Persons** | 994 | 1040 | 210 |
| **Pan Angles** | $\{-90, -75, -67.5, -45,$<br>$-22.5, -15, 0, +15,$<br>$+22.5, +45, +67.5, +90\}$ | $\{-45, -30,$<br>$-15, 0, +15,$<br>$+30, +45\}$ | $\{-67, -45, -22,$<br>$-15, 0, +15,$<br>$+22, +45, +67\}$ |
| **Tilt Angles** | 0 | $\{-45, 0, +45\}$ | 0 |
| **Age** | 19 years of birth | Y, M, O | Y, M, O |
| **Ethnicity** | 9 ethnic groups | Asian only | 5 ethnic groups |
| **# Images per subjects** | At least 5 by subject | $\sim 21$ | At least 5 by subject |

Table 5.1: Characteristics of the FERET, CAS-PEAL and our MIX datasets.

image feature $\mathbf{x}$ is performed using the labels distribution in $L(\mathbf{x};\mathcal{T})$ considering each characteristic distribution:

$$\text{label}(\mathbf{x};\mathcal{F})^c = \arg\max_{y^c} \frac{1}{|\mathcal{F}|} \sum_{\mathcal{T}\in\mathcal{F}} p(y^c|L(\mathbf{x};\mathcal{T})),\ \forall c = \{1,\ldots,C\}. \quad (5.18)$$

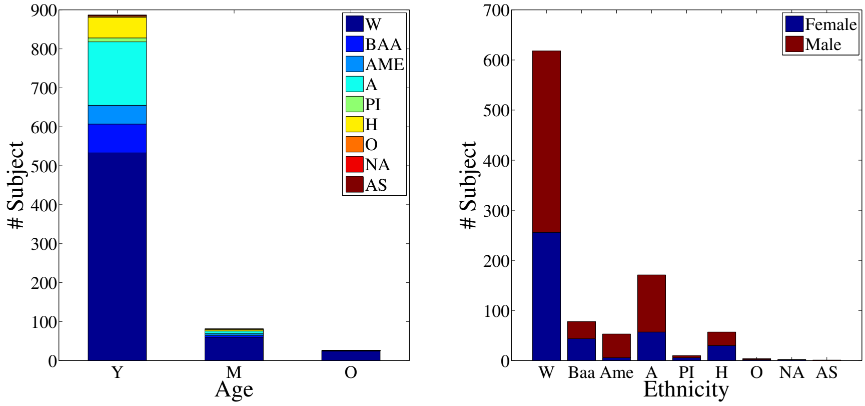
A detailed experimental analysis of our learning procedure is conducted in Section 5.3.2.

## 5.3 Experimental Results

With the same goal of the proposed algorithm, the authors in [36] investigate two different ways to estimate Gender, Age and Ethnicity using Kernel Canonical Correlation Analysis (KCCA) and PLS (Partial Least Squares). The work in [36] is hence being used as state-of-the-art baseline, comparing with the experiments of our approach. To properly evaluate the performance on our algorithm, we create a dataset using a subset of images from the FERET [72] and CAS-PEAL [32] datasets.

### 5.3.1 Datasets and Experimental Protocols

We evaluate our technique on a subset of the FERET [72] and CAS-PEAL-R1 [32] datasets (four examples of both dataset are illustrated in Figure 5.8). The main characteristics of these datasets are summarized in Table 5.1. We first detail each of these datasets, and then describe how we merge the two of them to create a combined dataset with a richer and more balanced set of soft biometrics to estimate.



(a) Ethnicity/Age (FERET)

(b) Gender/Ethnicity (FERET)

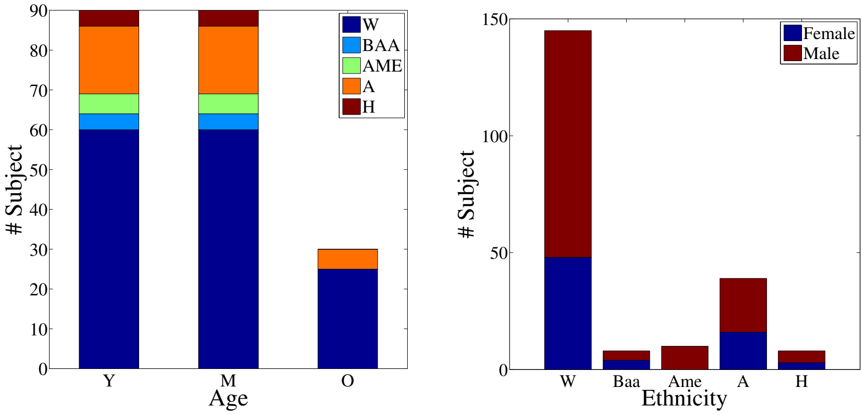
Figure 5.5: Distribution of characteristics in the FERET dataset. (a) Number of each ethnic groups within each Age class. (b) Gender distribution for each Ethnicity class.

### The FERET Dataset

The FERET dataset [72] is a benchmark for face-recognition algorithms. It is composed of images of 994 subjects (591 males and 403 females). It includes 11,338 images, annotated with 9 different ethnic groups, 12 different Pan angles, and year of birth.

### The CAS-PEAL-R1 Face Dataset

The complete CAS-PEAL dataset [32] is one of the largest datasets (99,594 images of 1,040 subjects) for evaluating Gender recognition, head pose estimation, and face recognition methods. The dataset is composed only of Chinese people (595 males and 445 females) imaged in different poses and with varying expressions, accessories, and lighting. The publicly available version, called CAS-PEAL-R1, contains 30,863 images of 1,040 subjects, with about twenty images of each person. Each image is acquired considering a combination of the Tilt and Pan angles reported in Table 5.1 and the Age is quantized into three classes Young (Y) from 10 to 44 years, Middle-Age (M) from 45 to 59 years and Old (O) above 60 years.



(a) Ethnicity/Age (MIX)

(b) Gender/Ethnicity (MIX)

Figure 5.6: Characteristic distribution in the MIX dataset. See text for definition of MIX dataset. (a) Ethnicity distribution in each Age class. (b) Distribution of genders in each Ethnicity class.

### A Dataset for Multiple Soft Biometric Estimation

Neither CAS-PEAL-R1 nor FERET alone are satisfactory for evaluation of multiple soft biometric estimation. For example there are only 4 “Old” subjects and 10 “Middle-Age” subjects in the CAS-PEAL-R1 dataset, and it also composed only of Asian subjects. In contrast, the FERET dataset contains very few “Old” subjects and 81 “Middle-Age” ones, and 9 different Ethnicity classes. As seen in Fig. 5.5(a), there are mostly “White/Young” people, and few “Middle-Age” and “Old” ones. Additionally, as shown in 5.5(b) Gender is not uniformly distributed within Ethnicity and some ethnicities have very few subjects.

We define a dataset that mixes images from the FERET and CAS-PEAL-R1 (which we call the “MIX” dataset in what follows). When joining the two datasets, we exclude images having a Pan angle of $\{-90, -75, -30, +30, +90\}$, because $-90$, $-75$ and $+90$ are present only in FERET and $\{-30, +30\}$ are present only in CAS-PEAL. We remove all the images with a nonzero Tilt angle, since FERET contains no images with nonzero tilt angles, and set to $-67$ and $67$ the Pan angles $-67.5$ and $67.5$, respectively. The images from 4 ethnic groups that have very few subjects were removed. We hence maintain the following 5 ethnicities: White (W), Black-or-African-American (BAA), Asian-Middle-Eastern (AME), Asian (A) and

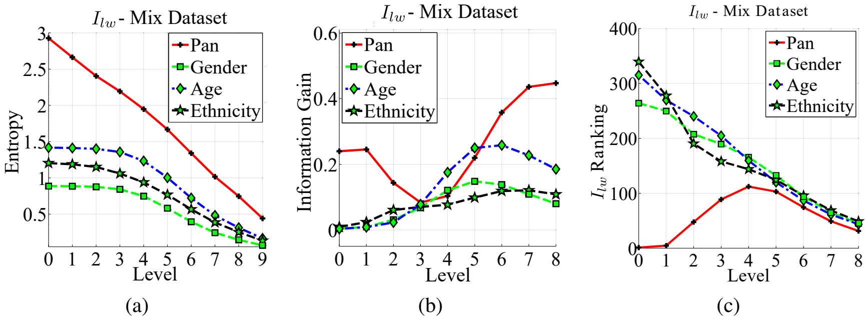


Figure 5.7: MORF performance on the MIX dataset. (a) Mean entropy per level for each characteristic. (b) Mean information gain of selected split per level. (c) Mean rank (in terms of information gain) of selected split per level.

Hispanic (H). Note that the authors of [36] addressed the imbalance in ethnicities by maintaining only the White and Asian classes. Finally, we project the Age labels from the FERET dataset into the set of Age classes of CAS-PEAL: Young, Middle and Old, as these three classes are the only Age information provided in CAS-PEAL.

In our dataset the majority number of subjects are “White” or “Asian”, and a significant part is composed by “Young” people as shown in Fig. 5.5(a). To avoid bias in classification results, we consider ten randomly drawn subsets of these subjects. Our random subset divisions consider the number of subjects for each Age class so that each subset contains 90 “Young” subjects, 90 “Middle-Age” subjects and 30 “Old” subjects, where the 90 “Young” are selected randomly from the mixed dataset, so that we have the same number of subjects for each Ethnicity class both for “Young” and “Middle-Age” people. Each split contains from 1,858 to 1,995 images, randomly divided in half of the subjects for each Ethnicity class to form the training and test sets.

### Experimental Protocol

Given the random nature of our method, and the ten different subsets, all results reported are averages over multiple runs of the algorithm (three trials for each split of the dataset). All trees used in these experiments have maximum depth of 9, which was found to work well in preliminary tests. Our random forest is composed of 200 trees. At each node of a tree, $T = 500$ parameterizations of the split function

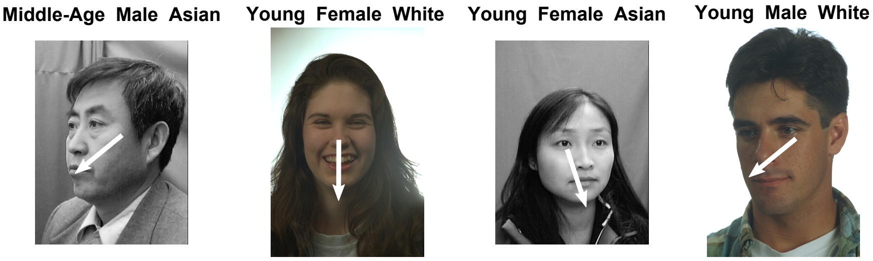


Figure 5.8: The four images show four different faces extracted from the CAS-PEAL dataset (the grey scale images) and the FERET dataset (the colored ones). The Age, Gender, Ethnicity and head pose (i.e. Pan angles) are estimate by MORF for each images.

are randomly generated and the best one (according to Eq. (5.17)) is selected.

### 5.3.2 Analysis of Multi-objective Learning

We evaluate Pan angles and soft biometrics estimation in terms of recognition accuracy, and only Pan angles and Age in terms of mean average error (MAE). While recognition accuracy and mean average error are commonly used for performance evaluation, we feel that – especially for soft biometrics in unbalanced datasets – precision and recall on individual biometrics is a better metric for evaluating performance. Moreover, we thus give precision/recall plots for all estimated soft biometrics.

In Figs. 5.7(a), 5.7(b) and 5.7(c) we show, respectively, the evolution of the entropy, the information gain, and the ranking (in terms of information gain) of each estimated characteristic as a function of the depth in the random trees comprising a MORF. The curves were computed from the average of all of the trees of all trials using our proposed information gain combination approach (Eq. (5.17)) with 200 trees. At the first levels of the trees, the Pan characteristic has the highest entropy (see Fig. 5.7(a)) and indeed as illustrated in Fig. 5.7(c) Pan is the highest ranked splitting characteristic and also results in the highest information gain (Fig. 5.7(b)). The early levels of each tree thus specialize in discriminating Pan angles and consequently implicitly condition the estimation of the other characteristics on Pan angle. Going deeper into the trees, the entropy of each soft biometric characteristic decreases gradually, and Age, Gender and Ethnicity begin to have more importance

| | Pan | | Age | | Gender | Ethnicity |
|---|---|---|---|---|---|---|
| Approach | Accuracy | MAE | Accuracy | MAE | Accuracy | Accuracy |
| MORF | **63.56%** | **10.65** | **58.89%** | **0.43** | 79.86% | **78.42%** |
| CCA | 31.27% | 19.29 | 47.91% | 0.58 | 76.28% | 43.71% |
| KCCA | 40.30% | 14.59 | 57.12% | **0.43** | **83.63%** | 40.04% |

Table 5.2: Comparison between the proposed approach and the state-of-the-art on Pan, Age, Gender and Ethnicity estimation.

as splitting characteristics. Trees learn to specialize first on Pan, then Gender and Age, and finally on Ethnicity.

### 5.3.3 Comparison with the State-of-the-art

In this section we compare the performance of our proposed MORF approach applied for Pan and soft biometric estimation, with those of the Canonical Correlation Analysis (CCA) and Kernel CCA (KCCA) approaches reported in [36]. These approaches have the same objective as MORF: to reliably and simultaneously estimate multiple face characteristics. They work by learning a subspace in which correlation between image features and desired characteristics is maximized, and then fitting a least squares prediction model from projected image features and characteristics. We evaluate baselines with standard linear CCA, and KCCA with a radial basis kernel.

We give accuracy and mean absolute error (MAE) performance of our method and both baselines in Table 5.2. We give Mean Average Precision (MAP) for all methods and for each label separately in Table 5.3. In Figs. 5.9, 5.11 and 5.10 the precision/recall curves of the approach compared with CCA and KCCA are plotted. They concern: Age (Figs. 5.9(a) and 5.9(b)), Gender (Figs. 5.10(a) and 5.10(b)), and Ethnicity estimation (Figs. 5.11(a) and 5.11(b)). Moreover, an example of qualitative results obtained with the proposed approach are shown in Figure 5.8.

#### Pan Estimation

We compare the Pan estimation performance of our algorithm with the state-of-the-art in the first two columns of Table 5.2 in terms of accuracy and MAE. The MAE measures how close predictions are to the expected outcomes. Our method outperforms the state-of-the-art baselines by a significant margin. The Pan charac-

| **Characteristic** | **Label** | **MORF** | **CCA** | **KCCA** |
|---|---|---|---|---|
| **Age** | Y | **54.66%** | 46.15% | 53.75% |
| | M | **59.20%** | 53.74% | 56.53% |
| | O | 19.64% | **23.85%** | 19.09% |
| **Gender** | F | 53.75% | 50.09% | **63.07%** |
| | M | 80.69% | 84.76% | **85.08%** |
| **Ethnicity** | W | 78.65% | 79.55% | **79.86%** |
| | BAA | 5.64% | **7.62%** | 6.55% |
| | AME | 6.33% | **7.62%** | 6.55% |
| | A | **57.73%** | 27.85% | 45.63% |
| | H | 4.08% | **4.94%** | 4.35% |

Table 5.3: Comparison between our approach and the state-of-the-art in terms of MAP on Age, Gender and Ethnicity estimation.

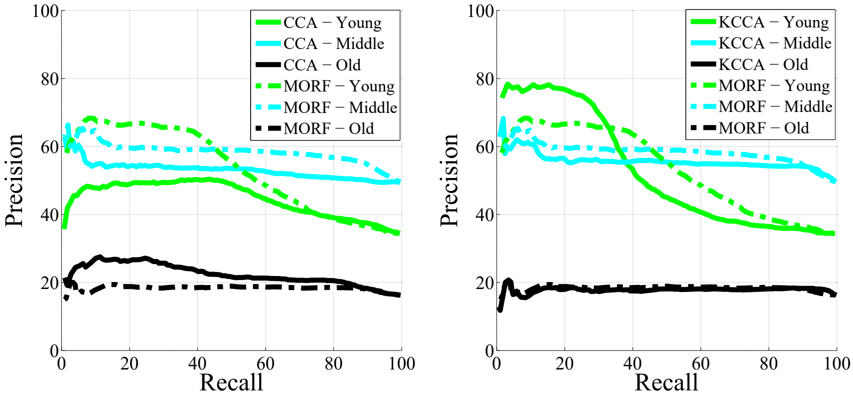
teristic has 9 different labels $(-67, -45, -22, -15, 0, +15, +22, +45, +67)$, this high number of labels may be an issue for the CCA and KCCA baselines.

### Age Estimation

We achieve similar or slightly higher performance with respect to the CCA and KCCA baselines. Accuracy and MAE are given in Table 5.2. We obtain an accuracy of $58.89\%$, which is $1.7\%$ higher than the KCCA baseline. In the first three rows of Table 5.3 we report MAP performance on Age estimation. We plot in Fig. 5.9 the precision/recall curves of our approach compared with CCA and KCCA baseline. The trend is similar for MORF and the two baselines, but our curves tend to stay above the CCA curves (except for the Old class) and a bit above the KCCA curves (except for the Young class for lower recall values).

### Gender estimation

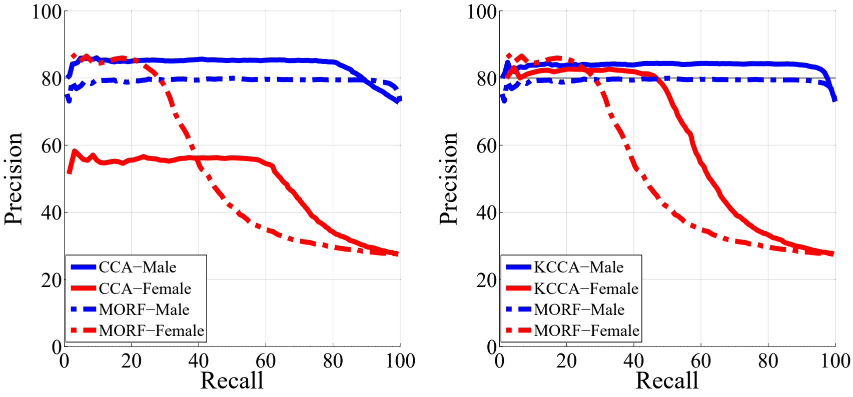
Our MORF approach gives lower performance that the KCCA baselines both in terms of accuracy and MAE as reported in Table 5.2. However, MORF outperforms the CCA baseline on these two metrics. From Table 5.3 where per class MAP are given and from the Precision/Recall curves in Fig. 5.10(a), we can see our method works better on female class with respect to the CCA, while in Fig. 5.10(b) we can see that the performance on the Female class is lower than KCCA. For all methods the performance on the female class is lower, this could be related to the fact that



(a) Age: CCA vs MORF

(b) Age: KCCA vs MORF

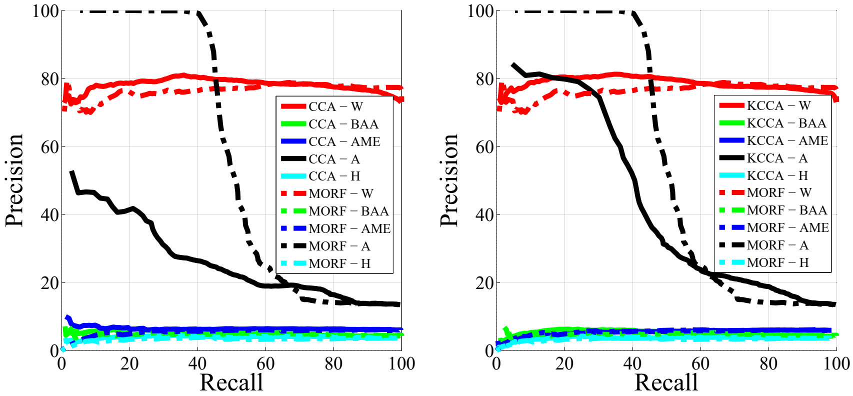
Figure 5.9: Precision/Recall of MORF for Age estimation. (a) Compared to CCA. (b) Compared to KCCA.



(a) Gender:CCA vs MORF

(b) Gender: KCCA vs MORF

Figure 5.10: Precision/Recall of MORF for Gender estimation. (a) Compared to CCA. (b) Compared to KCCA.



(a) Ethnicity: CCA vs MORF

(b) Ethnicity: KCCA vs MORF

Figure 5.11: Precision/Recall of MORF for Ethnicity estimation. (a) Compared to CCA. (b) Compared to KCCA.

only one third of the subjects in MIX dataset are females.

### Ethnicity Estimation

Our method significantly outperforms the baselines on Ethnicity estimation. MORF obtains an accuracy of $78.42\%$ while the CCA and KCCA baseline are around $40\%$, see Table 5.2. In terms of MAP, results in Table 5.3, the performances of our method are comparable with the baselines for most of labels except for the Asian class where our technique performs much better. This is confirmed by the precision/recall plot in Figs. 5.11(a) and 5.11(b), where the Asian curve of our approach is much higher.

## 5.4 Discussion

In this chapter we describe a technique for simultaneously estimating multiple facial characteristics. We propose a new normalized measure of multi-objective information gain that is used with our Multi-Objective Random Forests (MORF) framework for estimation of multiple characteristics from a single feature representation. On average, MORF outperforms subspace methods like KCCA and CCA for simultaneous estimation of multiple biometrics.

We feel that simultaneous estimation of characteristics is an interesting direction for future research. Ongoing work is focused on better features for soft biometric estimation, and on addressing problems where many characteristics (such as attributes) can be estimated for each sample. Methods to overcome the imbalance of the training sets are also something we plan to study.

# Chapter 6

# Conclusions

*Non puoi cambiare il passato,*
*ma puoi cambiare il futuro con quello che farai oggi.*

*You cannot change the past,*
*but you can change the future by what you will do today.*

- Anonymous

## 6.1 Concluding Remarks

The proposed methods have been defined to analyze video content and face images in order to automatically associate and track several target across camera in a video scene and jointly estimate multiple facial characteristics. Tracking is carried out automatically by interpreting a video stream using new detections, and we define two specific sub-problems: a data association method, and a long-term consistency technique to improve the performance of the defined end-to-end tracking system. For the first sub-problem we defined a sparse reconstruction method that associates new observations with the correct target trajectories. For the long-term consistency problem, a weighted version of the sparse reconstruction was proposed that focuses on the reduction of tracklet fragmentation and at the same time improves the performance of the system. Finally, the work on facial characteristic led to an approach based on the random forests that is able to simultaneously estimate multiple characteristics.

## 6.2 Summary of Contributions

The contributions of this thesis are related to two central topics to computer vision: visual tracking and face analysis. The research has led to the definition of new practical and efficient methods that could be implemented in video surveillance applications and video analytics software. Several aspects and issues have been covered in each topic that yield to performance results comparable to the state-of-the-art.

For visual tracking system we studied two different sub-problems: multi-target data association, and long-term consistency in a real scenarios, treated respectively in Chapters 3 and 4. In Chapter 3 we defined a data association algorithm that is robust to issues like occlusion and appearance change. We used an $\ell_1$-regularized optimization problem to define the proposed method, and studying only the pure data association problem we proposed a new discriminative model to robustly represent each identified target. The preliminary results show that the algorithm can be competitive with those ones used by state-of-the-art tracking methods.

The discriminative target model and the data association process proposed in Chapter 3 were then used to define an end-to-end multi-target tracking algorithm in Chapter 4. We pay special attention to the solution of the long-term consistency problem. We developed the model to take into account two different phases: a local data association and a global data association. The use of sparse reconstruction allows the tracking system to integrate observations of target appearance from the entire lifetime of each target. This significantly improves the local data association phase which appears more robust to occlusion and changing target appearance. Moreover, we have introduced a novel global data association phase that exploits the accumulated statistics of sparse local associations in order to merge tracklets likely to correspond to the same target. The developed technique obtained positive results comparable with the state-of-the-art, reducing the identity switches and tracklet fragmentation.

Finally, a technique for simultaneously estimating multiple facial characteristics was presented in Chapter 5. We proposed a new normalized measure of multi-objective information gain that was used jointly with the Multi-Objective Random Forests (MORF) framework to estimate Gender, Age, Ethnicity and head pose from a single feature representation. MORF obtains very good results for simultaneous estimation of multiple biometrics compared to KCCA- and CCA-based approaches.

## 6.3 Expected Impact of the Proposed Research

We believe that the proposed research can have a high impact in the fields of video surveillance applications, video analytics and face analysis, and in particular in the development of smart city applications. The two works on visual tracking (Chapters 3 and 4) propose a new way to handle the association and the tracking process of several targets. The used algorithm could be customized for distributed camera networks that use multiple cameras in wide areas for surveillance.

The proposed model adapts itself well in this situation because it can maintain several appearance aspects of each target in order to maintain robustness even if a target moves in and out camera views. Moreover, when applied in smart environments the algorithm can be used to profile target interests. A group interaction analysis system can apply the proposed system to analyze person interactions using target information collected by the tracker. Ongoing research is focused on integrating global and local data association into a single, continuous framework that obviates the need to run global association at arbitrary intervals during the tracking process.

Our facial characteristic estimation technique jointly estimates four different characteristics of face images and considers the face in its entirety. The proposed algorithm can be used in portable devices, because the experiments we carried out proved that the obtained results are comparable with the state-of-the-art at low resolutions. For smart environments, our algorithm can be used to estimate relevant characteristics and to propose customized multimedia contents or to adapt advertising displayed on nearby screens. The simultaneous estimation of characteristics is an interesting direction for future research.

Current work is focused on better features for soft biometric estimation, considering the estimation of many facial attributes. The methods defined for both topics are very interconnected, and could be used to develop a profiling system for the subject interest through the tracking process in smart environments, for instance in shopping mall to propose commercial content-based considering relevant characteristics or to evaluate how long a person focuses on a particular store. In an hypothetical smart city environment a method like this could be adopted for tourism as well as for security purposes.

# Appendix A

# Publications

This research activity has led to three publications in international conferences and in a journal. These are listed below.

## International Conferences

- Andrew D. Bagdanov, Alberto Del Bimbo, **Dario Di Fina**, Svebor Karaman, Giuseppe Lisanti, Iacopo Masi, "**Multi-Target Data Association using Sparse Reconstruction**", in Proceedings of *International Conference on Image Analysis and Processing* (ICIAP), Naples, Italy, 2013.

- **Dario Di Fina**, Svebor Karaman, Andrew D. Bagdanov and Alberto Del Bimbo, "**MORF: Multi-Objective Random Forests for Face Characteristic Estimation**", in Proceedings of *International Conference on Advanced Video- and Signal-based Surveillance* (AVSS), Karlsruhe, Germany, 2015.

## Submitted Journal

- **Dario Di Fina**, Giuseppe Lisanti, Svebor Karaman, Andrew D. Bagdanov and Alberto Del Bimbo, "**Multi-Target Tracking using Weighted Sparse Reconstruction**", *Submitted to Pattern Analysis and Applications*, 2015.

The author's bibliometric indices are the following: $H$-index = 1, total number of citations = 1 (source: Google Scholar on Ottobre 28, 2015).

# Bibliography

[1] Pets 2009 benchmark data. [Online]. Available: http://www.cvg.reading.ac.uk/PETS2009/ 1, 23, 32, 50

[2] A. Andriyenko and K. Schindler, "Multi-target tracking by continuous energy minimization," in *Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, June 2011, pp. 1265–1272. 11, 22

[3] S.-H. Bae and K.-J. Yoon, "Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning," in *Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, 2014, pp. 1218–1225. 5, 38, 46, 50, 52, 53

[4] A. D. Bagdanov, A. Del Bimbo, D. Di Fina, S. Karaman, G. Lisanti, and I. Masi, "Multi-target data association using sparse reconstruction," in *Proceedings of International Conference on Image Analysis and Processing (ICIAP).* Springer, 2013, vol. 8157, pp. 239–248. 11, 21

[5] Y. Bar-Shalom and E. Tse, "Tracking in a cluttered environment with probabilistic data association," *Automatica*, vol. 11, no. 5, pp. 451 – 460, Elsevier, 1975. 12, 13

[6] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, 2011, pp. 3457–3464. 1

[7] ——, "Unsupervised learning of a scene-specific coarse gaze estimator," in *Proceedings of International Conference on Computer Vision (ICCV).* IEEE, 2011, pp. 2344–2351. 18, 19

[8] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *Journal Image Video Processing*, vol. 2008, pp. 1:1–1:10, 2008. [Online]. Available: http://dx.doi.org/10.1155/2008/246309 32, 33

[9] M. Bredereck, X. Jiang, M. Korner, and J. Denzler, "Data association for multi-object tracking-by-detection in multi-camera networks," in *Proceedings of International Conference on Distributed Smart Cameras (ICDSC).* IEEE, 2012, pp. 1–6. 12

[10] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, “Robust tracking-by-detection using a detector confidence particle filter,” in *Proceedings of International Conference on Computer Vision (ICCV)*. IEEE, 2009, pp. 1515–1522. 12

[11] ——, “Online multiperson tracking-by-detection from a single, uncalibrated camera,” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1820–1833, 2011. 32, 33, 34

[12] E. J. Candes, M. B. Wakin, and S. P. Boyd, “Enhancing sparsity by reweighted l1 minimization,” *Journal of Fourier Analysis and Applications*, vol. 14, no. 5-6, pp. 877–905, Springer, 2008. 46

[13] L. Chen, W. Wang, G. Panin, and A. Knoll, “Hierarchical grid-based multi-people tracking-by-detection with global optimization,” *IEEE, Transactions on Image Processing*, vol. 24, no. 11, pp. 4197–4212, 2015. 16

[14] X. Chen and H. Ishwaran, “Random forests for genomic data analysis,” *Genomics*, vol. 99, no. 6, pp. 323–329, Elsevier, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0888754312000626 60

[15] A. Ciolini, L. Seidenari, S. Karaman, and A. Del Bimbo, “Efficient hough forest object detection for low-power devices,” in *Proceedings of International Conference on Multimedia Expo Workshops (ICMEW)*. IEEE, 2015, pp. 1–6. 60

[16] I. J. Cox, “A review of statistical data association techniques for motion correspondence,” *International Journal of Computer Vision*, vol. 10, pp. 53–66, Springer,1993. 12, 13, 14, 22

[17] L. J. Cox and S. L. Hingorani, “An efficient implementation of reid’s multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 2, pp. 138 –150, 1996. 14

[18] A. Criminisi, J. Shotton, and E. Konukoglu, “Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning,” *Foundations and Trends® in Computer Graphics and Vision*, vol. 7, no. 2, pp. 81–227, 2012. 56, 59, 60

[19] M. Cristani, R. Raghavendra, A. Del Bue, and V. Murino, “Human behavior analysis in video surveillance: A social signal processing perspective,” *Neurocomputing*, vol. 100, pp. 86–97, Elsevier, 2013. 1

[20] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, IEEE, 2005, pp. 886–893. 17, 41, 61

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 248–255. 60

[22] D. Di Fina, S. Karaman, A. D. Bagdanov, and A. Del Bimbo, “MORF: Multi-objective random forests for face characteristic estimation,” in *International Conference on Advanced Video- and Signal-based Surveillance (AVSS).* IEEE, 2015, pp. 1–6. 55

[23] A. Doucet, N. De Freitas, K. Murphy, and S. Russell, “Rao-blackwellised particle filtering for dynamic bayesian networks,” in *Proceedings of Conference on Uncertainty in Artificial Intelligence (UAI).* Morgan Kaufmann Publishers Inc., 2000, pp. 176–183. [Online]. Available: http://dl.acm.org/citation.cfm?id=647234.720075 14

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

[25] A. Ess, B. Leibe, K. Schindler, and L. V. Gool, “A mobile vision system for robust multi-person tracking,” in *Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, 2008, pp. 1–8. 50, 52

[26] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (VOC) challenge,” *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, Jun. Springer, 2010. 29

[27] G. Fanelli, J. Gall, and L. Van Gool, “Real time head pose estimation with random regression forests,” in *Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2011, pp. 617–624. 9, 60

[28] T. E. Fortmann and Y. Bar-Shalom, *Tracking and Data Association.* Boston: Academic-Press, 1988. 13, 14

[29] T. E. Fortmann, Y. Bar-Shalom, and M. Scheffe, “Sonar tracking of multiple targets using joint probabilistic data association,” *Journal of Oceanic Engineering*, vol. 8, no. 3, pp. 807–812, IEEE, 1983. 12, 14

[30] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky, “Hough forests for object detection, tracking, and action recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2188–2202, 2011. 60

[31] A. Gallagher and T. Chen, “Understanding images of groups of people,” in *Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, 2009, pp. 256–263. 18

[32] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao, “The CAS-PEAL large-scale chinese face database and baseline evaluations,” *Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans.*, vol. 38, no. 1, pp. 149–161, IEEE, 2008. 65, 66

[33] W. Ge, R. T. Collins, and R. B. Ruback, “Vision-based analysis of small groups in pedestrian crowds,” *Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 5, pp. 1003–1016, IEEE, 2012. 1

[34] N. Gourier, J. Maisonnasse, D. Hall, and J. L. Crowley, “Head pose estimation on low resolution images,” in *Proceedings of International Evaluation Conference on Classification of Events, Activities and Relationships (CLEAR).* Springer-Verlag, 2006, pp. 270–280. [Online]. Available: http://dl.acm.org/citation.cfm?id=1759639.1759670 16, 55

[35] A. Gunay and V. V. Nabiyev, “Automatic age classification with LBP,” in *Proceedings of International Symposium on Computer and Information Sciences (ISCIS).* IEEE, 2008, pp. 1–4. 17

[36] G. Guo and G. Mu, “A framework for joint estimation of age, gender and ethnicity on a large database,” *Image and Vision Computing*, vol. 32, no. 10, pp. 761 – 770, Elsevier, 2014. 18, 65, 68, 70

[37] G. Guo, C. R. Dyer, Y. Fu, and T. S. Huang, “Is gender recognition affected by age?” in *Proceedings of International Conference on Computer Vision Workshops (ICCV Workshops).* IEEE, 2009, pp. 2032–2039. 17

[38] M. A. Haj, J. Gonzalez, and L. S. Davis, “On partial least squares in head pose estimation: How to simultaneously deal with misalignment,” in *Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, 2012, pp. 2602–2609. 19

[39] J. Han, E. J. Pauwels, P. M. De Zeeuw, and P. H. N. De With, “Employing a RGB-D sensor for real-time tracking of humans across multiple re-entries in a smart environment,” *Transactions on Consumer Electronics*, vol. 58, no. 2, pp. 255–263, IEEE, 2012. 2

[40] A. Hapfelmeier and K. Ulm, “A new variable selection approach using random forests,” *Transactions on Computational Statistics & Data Analysis*, vol. 60, pp. 50–69, Elsevier, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167947312003490 60

[41] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, “The elements of statistical learning: data mining, inference and prediction,” *The Mathematical Intelligencer*, vol. 27, no. 2, pp. 83–85, Springer, 2005. 23, 24

[42] H. T. Ho and R. Chellappa, “Automatic head pose estimation using randomly projected dense sift descriptors,” in *International Conference on Image Processing (ICIP).* IEEE, 2012, pp. 153–156. 18

[43] A. Hoogs and A. G. A. Perera, “Video activity recognition in the real world,” in *Proceedings of National Conference on Artificial Intelligence (AAAI).* AAAI Press, 2008, pp. 1551–1554. [Online]. Available: http://dl.acm.org/citation.cfm?id=1620270.1620330 1

[44] C. Huang, X. Ding, and C. Fang, “Head pose estimation based on random forests for multiclass classification,” in *Proceedings of International Conference onPattern Recognition (ICPR).* IEEE, 2010, pp. 934–937. 9, 16, 55, 60

[45] 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, Tech. Rep., 2007. 18

[46] X. Jia, H. Lu, and M.-H. Yang, “Visual tracking via adaptive structural local sparse appearance model,” in *Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, 2012, pp. 1822–1829. 24

[47] H. Jiang, J. Wang, Y. Gong, N. Rong, Z. Chai, and N. Zheng, “Online multi-target tracking with unified handling of complex scenarios,” *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3464–3477, 2015. 13

[48] Z. Jiang, Z. Lin, and L. S. Davis, “Learning a discriminative dictionary for sparse coding via label consistent K-SVD,” in *Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, 2011, pp. 1697–1704. 23

[49] S. Karaman, A. D. Bagdanov, G. D’Amico, L. Landucci, A. Ferracani, D. Pezzatini, and A. Del Bimbo, “Passive profiling and natural interaction metaphors for personalized multimedia museum experiences,” in *Proceedings of International On Image Analysis and Processing - New Trends in Image Analysis and Processing (ICIAP).* Springer, 2013, vol. 8158, pp. 247–256. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-41190-8_27 11, 18

[50] S. Karaman, G. Lisanti, A. D. Bagdanov, and A. Del Bimbo, “Leveraging local neighborhood topology for large scale person re-identification,” *Pattern Recognition*, vol. 47, pp. 3767–3778, 2014. 50

[51] J. Kokkala and S. Sarkka, “Combining particle MCMC with rao-blackwellized monte carlo data association for parameter estimation in multiple target tracking,” *Digital Signal Processing*, vol. 47, pp. 84 – 95, Elsevier, 2015. 13, 22

[52] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, “Attribute and simile classifiers for face verification,” in *Proceedings of International Conference on Computer Vision (ICCV).* IEEE, 2009, pp. 365–372. 18

[53] C.-H. Kuo, C. Huang, and R. Nevatia, “Multi-target tracking by on-line learned discriminative appearance models,” in *Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, 2010, pp. 685–692. 14, 28

[54] H.-C. Lian and B.-L. Lu, “Multi-view gender classification using local binary patterns and support vector machines,” in *Proceedings of International Conference on Advnaces in Neural Networks (ISNN)*, vol. 3972, Springer, 2006, pp. 202–209. [Online]. Available: http://dx.doi.org/10.1007/11760023_30 16, 55

[55] G. Lisanti, I. Masi, A. Bagdanov, and A. Del Bimbo, “Person re-identification by iterative re-weighted sparse ranking,” *Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 8, pp. 1629–1642, IEEE, 2014. 38, 40, 46, 50

[56] G. Lisanti, I. Masi, and A. Del Bimbo, "Matching people across camera views using kernel canonical correlation analysis," in *Proceedings of the International Conference on Distributed Smart Cameras (CDSC)*. ACM, 2014, p. 10. 50

[57] R. Liu, Y. Chen, X. Zhu, and K. Hou, "Image classification using label constrained sparse coding," *Multimedia Tools and Applications*, vol. 2015, pp. 1–15, Springer, 2015. [Online]. Available: http://dx.doi.org/10.1007/s11042-015-2626-1 23

[58] W. Liu, D. Tao, J. Cheng, and Y. Tang, "Multiview hessian discriminative sparse coding for image annotation," *Computer Vision and Image Understanding*, vol. 118, pp. 50 – 60, Elsevier, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077314213001550 23

[59] B. Ma, W. Zhang, S. Shan, X. Chen, and W. Gao, "Robust head pose estimation using LGBP," in *Proceedings of International Conference on Pattern Recognition (ICPR)*. IEEE, 2006, pp. 512–515. [Online]. Available: http://dx.doi.org/10.1109/ICPR.2006.1006 18

[60] Y. Ma, J. Liu, X. Yang, Y. Liu, and N. Zheng, "Double layer multiple task learning for age estimation with insufficient training samples," *Neurocomputing*, vol. 147, pp. 380–386, Elsevier, 2015. 17

[61] Z. Ma and J. Wan, "Survey of data association of moving objects tracking in video sensors network," in *Proceedings of International Conference on Electronic Measurement & Instruments (ICEMI)*. IEEE, 2009, pp. 4–250–4–254. 13, 22

[62] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," in *Proceedings of Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2008, pp. 1033–1040. [Online]. Available: http://papers.nips.cc/paper/3448-supervised-dictionary-learning.pdf 23

[63] I. Masi, "From motion to faces: 3d-assisted automatic analysis of people," Ph.D. dissertation, University of Florence, 2013. 8

[64] I. Masi, G. Lisanti, A. D. Bagdanov, P. Pala, and A. Del Bimbo, "Using 3d models to recognize 2d faces in the wild," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*. IEEE, 2013, pp. 775–780. 8

[65] X. Mei and H. Ling, "Robust visual tracking using l-1 minimization," in *Proceedings of International Conference on Computer Vision (ICCV)*. IEEE, 2009, pp. 1436 –1443. 23, 24

[66] S. Mittal, T. Prasad, S. Saurabh, X. Fan, and H. Shin, "Pedestrian detection and tracking using deformable part models and kalman filtering," in *International SoC Design Conference (ISOCC)*, Nov IEEE, 2012, pp. 324–327. 5, 12, 15

[67] K. P. Murphy, "Switching kalman filters," Citeseer, Tech. Rep., 1998. 13

[68] C. B. Ng, Y. H. Tay, and B. Goi, "Vision-based human gender recognition: A survey," *arXiv:1204.1611*, vol. abs/1204.1611, 2012. 17

[69] S. Oh, S. Russell, and S. Sastry, “Markov chain monte carlo data association for general multiple-target tracking problems,” in *Proceedings of Conference on Decision and Control (CDC)*, vol. 1. IEEE, 2004, pp. 735 –742. 14

[70] B. Panti, P. Monteiro, F. Pereira, and J. Ascenso, “Descriptor-based adaptive tracking-by-detection for visual sensor networks,” in *International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2015, pp. 1–6. 12

[71] A. Pentland and T. Choudhury, “Face recognition for smart environments,” *Computer*, vol. 33, no. 2, pp. 50–55, IEEE, 2000. 2

[72] P. J. Phillips, H. Moon, S. Rizvi, P. J. Rauss *et al.*, “The feret evaluation methodology for face-recognition algorithms,” *Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, IEEE, 2000. 17, 65, 66

[73] R. Pichevar, H. Najaf-Zadeh, L. Thibault, and H. Lahdili, “Auditory-inspired sparse representation of audio signals,” *Transactions on Speech Communication*, vol. 53, no. 5, pp. 643–657, May Elsevier, 2011. 24

[74] R. Ramamoorthi, “Sparse reconstruction of visual appearance for computer graphics and vision,” EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2011-86, 2011. 23

[75] N. Ramanathan, R. Chellappa, S. Biswas *et al.*, “Age progression in human faces: A survey,” *Journal of Visual Languages and Computing*, vol. 15, pp. 3349 – 3361, 2009. 17

[76] K. Ramesha, K. B. Raja, K. R. Venugopal, and L. M. Patnaik, “Feature extraction based face recognition, gender and age classification,” *Transactions on International Journal of Computer Science and Engineering (IJCSE)*, vol. 2, no. 1, pp. 14–23, 2010. 17

[77] K. Ricanek Jr. and T. Tesafaye, “MORPH: A longitudinal image database of normal adult age-progression,” in *Proceedings of International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2006, pp. 341–345. [Online]. Available: http://dx.doi.org/10.1109/FGR.2006.78 17

[78] M. Ristin, M. Guillaumin, J. Gall, and L. Van Gool, “Incremental learning of random forests for large-scale image classification,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2015. 60

[79] N. Robertson and I. Reid, “Estimating gaze direction from low-resolution faces in video,” in *Proceedings of European Conference on Computer Vision (ECCV)*. Springer, 2006, pp. 402–415. [Online]. Available: http://dx.doi.org/10.1007/11744047_31 1

[80] S. Santhoshkumar, S. Karthikeyan, and B. Manjunath, “Robust multiple object tracking by detection with interacting markov chain monte carlo,” in *Proceedings of International Conference on Image Processing (ICIP)*. IEEE, 2013, pp. 2953–2957. 15

[81] A. Schumann, M. Bauml, and R. Stiefelhagen, "Person tracking-by-detection with efficient selection of part-detectors," in *Proceedings of International Conference on Advanced Video and Signal Based Surveillance (AVSS).* IEEE, 2013, pp. 43–50. 16

[82] A. V. Segal and I. Reid, "Latent data association: Bayesian model selection for multi-target tracking," in *Proceedings of International Conference on Computer Vision (ICCV).* IEEE, 2013, pp. 2904–2911. 13

[83] X. Shi, H. Ling, W. Hu, C. Yuan, and J. Xing, "Multi-target tracking with motion context in tensor power iteration," in *Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2014, pp. 3518–3525. 14

[84] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah, "Part-based multiple-person tracking with partial occlusion handling," in *Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, June IEEE, 2012, pp. 1815–1821. 5, 12, 15

[85] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, pp. 267–288, 1996. 23, 24

[86] J. Tu, Y. Fu, Y. Hu, and T. Huang, "Evaluation of head pose estimation for studio data," in *Proceedings of International Evaluation Conference on Classification of Events, Activities and Relationships (CLEARW).* Springer-Verlag, 2006, pp. 281–290. [Online]. Available: http://dl.acm.org/citation.cfm?id=1759639.1759671 18

[87] M. Voit, K. Nickel, and R. Stiefelhagen, "Neural network-based head pose estimation and multi-view fusion," in *Proceedings of International Evaluation Conference on Classification of Events, Activities and Relationships (CLEAR).* Springer-Verlag, 2007, pp. 291–298. [Online]. Available: http://dl.acm.org/citation.cfm?id=1759639.1759672 18

[88] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 31, no. 2, pp. 210–227, IEEE, 2009. 23, 25

[89] T.-X. Wu and B.-L. Lu, "Multi-view gender classification using hierarchical classifiers structure," in *Proceedings of International Conference on Neural Information Processing: Models and Applications (ICONIP)*, vol. 6444, no. 2. Springer, 2010, pp. 625–632. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-17534-3_77 7

[90] Z. Wu, A. Thangali, S. Sclaroff, and M. Betke, "Coupling detection and data association for multiple object tracking," in *Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, 2012, pp. 1948–1955. 25

[91] B. Xia, H. Sun, and B.-L. Lu, "Multi-view gender classification based on local gabor binary mapping pattern and support vector machines," in *Proceedings of International Joint Conference on Neural Networks (IJCNN).* IEEE, 2008, pp. 3388–3395. 17

[92] B. Yang and R. Nevatia, “Multi-target tracking by online learning of non-linear motion patterns and robust appearance models,” in *Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2012, pp. 1918–1925. 14, 16

[93] ——, “Online learned discriminative part-based appearance models for multi-human tracking,” in *Proceedings of European Conference on Computer Vision (ECCV)*, Springer, 2012, pp. 484–498. 32, 33, 34

[94] ——, “Multi-target tracking by online learning a CRF model of appearance and motion patterns,” *International Journal of Computer Vision*, vol. 107, no. 2, pp. 203–217, Springer, 2014. 5, 15, 16, 39, 40, 50, 51, 52, 53

[95] M. Yang, L. Zhang, J. Yang, and D. Zhang, “Robust sparse coding for face recognition,” in *Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, June 2011, pp. 625–632. 23

[96] Z. Yang and H. Ai, “Demographic classification with local binary patterns,” in *Proceedings of International Conference on Advances in Biometrics (ICB).* Springer Berlin Heidelberg, 2007, vol. 4642, pp. 464–473. [Online]. Available: http://dl.acm.org/citation.cfm?id=2391659.2391713 17

[97] M. Yaqub, M. K. Javaid, C. Cooper, and J. A. Noble, “Investigation of the role of feature selection and weighted voting in random forests for 3-d volumetric segmentation,” *Transactions on Medical Imaging*, vol. 33, no. 2, pp. 258–271, IEEE, 2014. 60

[98] A. Yilmaz, O. Javed, and M. Shah, “Object tracking: A survey,” *Computing Surveys (CSUR)*, vol. 38, no. 4, p. 13, ACM, 2006. [Online]. Available: http://doi.acm.org/10.1145/1177352.1177355 12, 14

[99] J. Ylioinas, A. Hadid, and M. Pietikainen, “Age classification in unconstrained conditions using LBP variants,” in *Proceedings of International Conference on Pattern Recognition (ICPR).* IEEE, 2012, pp. 1257–1260. 17

[100] J. Yu, Y. Rui, and D. Tao, “Click prediction for web image reranking using multimodal sparse coding,” *Transactions on Image Processing*, vol. 23, no. 5, pp. 2019–2032, IEEE, 2014. 23, 24

[101] L. Zhang, Y. Li, and R. Nevatia, “Global data association for multi-object tracking using network flows,” in *Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, 2008, pp. 1–8. 13

[102] S. Zhang, J. Wang, Z. Wang, Y. Gong, and Y. Liu, “Multi-target tracking by learning local-to-global trajectory models,” *Pattern Recognition*, vol. 48, no. 2, pp. 580 – 590, Elsevier, 2015. 16

[103] J. Zheng and B.-L. Lu, “A support vector machine classifier with automatic confidence and its application to gender classification,” *Neurocomputing*, vol. 74,

no. 11, pp. 1926–1935, Elsevier, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231211000464 7, 16, 17, 55

[104] W. Zhong, H. Lu, and M.-H. Yang, "Robust object tracking via sparse collaborative appearance model," *Transactions on Image Processing*, vol. 23, no. 5, pp. 2356–2368, IEEE, 2014. 25, 38

[105] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2012, pp. 2879–2886. 7