Università degli Studi di Firenze
Corso di Dottorato in Informatica, Sistemi e Telecomunicazioni
Media Integration and Communication Center (MICC)
ING-INF/05

From motion to faces: 3D-assisted automatic analysis of people

Candidate
Iacopo Masi

Supervisors
Prof. Alberto Del Bimbo

Dr. Andrew D. Bagdanov

Dr. Federico Pernici

PhD Coordinator
Prof. Luigi Chisci

ciclo XXVI, 2011-2013

Università degli Studi di Firenze, Media Integration and Communication Center (MICC).

Alle persone a cui questa tesi
ha sottratto, forse, troppo tempo:
la mia compagna, i miei genitori.
Ai miei nonni che non ho mai conosciuto.

"If you live each day as if it was your last,
some day you'll most certainly be right."

Acknowledgments

I would like to acknowledge the effort and input of my supervisor, Professor Alberto Del Bimbo. Thank you. I also would like to thank Dr. Andrew D. Bagdanov and Dr. Federico Pernici, who guided me and collaborated on the main parts of my research work and who taught me two different but valid ways to approach the research. They have been also good friends and colleagues and not detached tutors, thank you. My thanks go to my colleagues of the Media Integration and Communication Center (MICC) who were of great help during my research. I would like to underline that the MICC is one of the best places, if not the best, to perform research and applied research in Italy. In particular, my special thanks go to Dr. Giuseppe Lisanti, which substantially contributed to this thesis. I would like also to thank Dr. Lorenzo Seidenari, Dr. Svebor Karaman and at last but not least, Dr. Marco Bertini, Dr. Lamberto Ballan, Dr. Giuseppe Serra, Andrea Ferracani, Daniele Pezzatini, Dr. Lea Landucci, Tiberio Uricchio, Federico Bartoli, Dario Di Fina, Claudio Baecchi, Francesco Turchini, Simone Ercoli, Maxime Devanne, Joël Dumoulin, Professor Pala, Professor Berretti and the other fellows that I met during my staying at MICC (that I cannot list here). I would like to thank also Professor Dr. Gérard Medioni, who gave me the possibility to do an internship in his research group, USC IRIS Computer Vision Lab, in Los Angeles. This has greatly advanced my career pursuits. I thank all the people that collaborated with me there: now I know what means to be a USC Trojan! Finally my thanks go also to Ing. Andrea Grifoni for the collaboration with Thales Italy S.p.A. and for his friendship.

Contents

Chapter 1

Introduction and Motivation

Observing people is one of the most important topic across all disciplines. Putting aside for the moment the technological fields, which is the aspect considered by this thesis, also social and psychological sciences have remarked on the importance of people and their interactions [143]. Among all the sciences, if we consider technological fields [47], there are plenty of applications of "person analysis". In general we can define the term person analysis as

> non-intrusive methods and processes that extract and pool coherent statistics per subject and then additionally infer a pattern that represents a behavior connected with a group of people or with the single subject.

This thesis focuses on ways of observing people which are performed considering three factors. Firstly, the analysis should be performed *automatically* without relying on consistent and constant human-interaction. Considering this, we can claim that this thesis aims to program machines that automatically observe people and then, of course, humans must only inspect the final results. This automation is reached by observing people at different levels using regular visual sensors such as fixed or moving cameras. Secondly, a more subtle aspect is that this machine-vision analysis should be assisted by 3D reasoning. Even if regular passive sensors such as CCTVs (Closed-Circuit Televisions) passively observe through media like images or video, we argue that the use of 3D information could enhance the performance of this automation. Finally, the analysis is performed interpreting visual content

with a computer, i.e. pixels in the image; and the science that has the role of understanding what is happening in imagery is computer vision.

Some interesting use cases of person analysis can be given if we consider our current daily life: for example think of what a security company could do if it were possible to automatically monitor a wide area without employing watchmen or anti-theft devices; or if it were possible to prevent theft before it happens by simply observing people through regular CCTVs. If it is impossible to prevent crimes, "analysis of people" could also be used to passively record interesting parts of videos for post-mortem analysis of events, in case someday this important data might be useful. Note that this automatic approach is not concerned with *passively* saving all the data, which is already performed by NVRs (Network Video Recorders), but with automatically keeping only the *relevant* ones.

However, "analysis of people" is not just a matter of security: it could also enhance our daily life when we shop. Imagine a supermarket that, respecting the privacy of users[1], is able to accumulate coherent statistics per subject about the places and items that have received interest. This could be in principle used to improve the way we shop in a two-fold manner: (1) the seller could have a statistics about person behavior in buying things, the physical places that they often visit, etc. (2) it could change the relationship between consumer and shop with respect to how the customer shops. We can imagine a future system suggests products to the customer, not only on his PC or smartphone, but directly *in place* in the supermarket. This latter use case provides a parallelism about how we are nowadays constantly monitored and observed by internet companies through the web in order to better improve their advertisement algorithms and, hopefully, the customer approach to their products.

In a near future, we can imagine that this non-intrusive way of observing people could also be extended not only to the virtual world of the web, but also to our physical daily life, observed not through protocols or clicks but using visual sensors such as visual cameras, IP and robotic cameras, depth sensors and robots that walk among us.

[1] This important topic and issues of privacy are beyond the scope of this thesis.

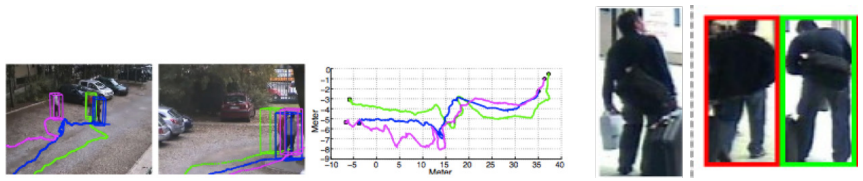
1.1 Objectives of this Thesis

What is the practical meaning of visual analysis of people? To have an idea, we present some questions that a machine must deal with when it is subject to analyze a video stream or still images. The following questions encompass some notoriously difficult tasks if we think that they need to be performed automatically by a machine: are there people present in the image? If yes, where are they located? Can you extract their trajectories in an absolute reference system? Are you able to guide the sensor in order to follow them and to focus on biometrics details such as the face? If I give a person image from camera A, can you please tell me who is this person when viewed from camera B? Can you recognize a face captured in the wild from a surveillance camera over a large set of candidates? How long does it take to recognize the person?

Considering this tasks, the objectives of this thesis are the following: we present a number of principled methods to answer these questions. In general, we analyze the motion of people in dynamic scenes, and considering that the sensors we employ is moving one, we can also think to steer the sensor in order to follow the people; then we propose a way to coarsely retrieve the path of a person by providing a method to link its path through multiple non-overlapping fixed cameras; we consider two cameras in this thesis, but extension to multiple cameras is straightforward; then, focusing faces the most non-intrusive biometric of all, we present two novel methods to recognize faces in the wild. These break the proposed thesis in two parts:

1. the first part concerns "motion analysis", and deals with the topic of person tracking and re-identification with the aim of extracting the motion and moving the sensor towards important details. An illustration of these topics is shown in Fig. 1.1.
2. the second part is about "face analysis" in term of face recognition and 3D face modeling and assumes that the method proposed in the first part has succeeded in the task of focusing on the face. An illustration of these issues are shown in Fig. 1.2.

This gives a rough idea of the topics and the organization of this thesis. Regarding the latter, more details will be given in the next Section.



(a) Trajectory Extraction (b) Person Re-identification

Figure 1.1: Analysis of people through motion encompasses trajectory extraction and person re-identification.

(a) Face Logs (b) 3D Face Pose Estimation

Figure 1.2: Analysis of people through faces deals with face logging, face recognition and face pose estimation.

1.2 Organization of this Thesis

As reported in the previous section, this thesis is organized in two parts: the first part concerns with the topic of motion analysis while the second part deals with face analysis in term of face modeling and recognition. In the all the chapters we first introduce and define the problem we are addressing and then we proceed to review the literature for that topic. In Fig. 1.3 we show a schematic viewpoint of the organization of this thesis: from the input that is a generic person, to the output of this work, in form of trajectories or an identity derived from a recognized face.

Moreover, in Table 1.1 we summarize the organization of this thesis in terms of how *3D data* assist the method, specifying if the 3D is inferred from the 2D imagery or it it is provided *a priori*.

Chapter	#	2D data	3D data	3D source
Part I: Motion Analysis				
Motion Analysis from a PTZ Camera	2	✓	✓	Inferred from 2D
Matching People across Views	3	✓	✗	✗
Part II: Face Analysis				
Logging of Face Imagery	4	✓	✗	✗
2D/3D Florence Faces Dataset (DB)	5	✓	✓	Manual acquisition
Matching Face Imagery with 3D Textured Models	6	✓	✓	Prior from DB
3D Assisted Pose Independent Face Recognition	7	✓	✓	Prior from DB

Table 1.1: The use of 2D/3D information in the chapters of this thesis along with the procedure that gives the 3D data.

Motion analysis

The topic of trajectory extraction and the task of focusing on biometric details is covered in Chapter 2. This is addressed in term of joint camera calibration and multi-person tracking. A key task in person analysis is the extraction of meaningful trajectories of people that are moving in a wide area while the sensor is zooming on biometric details. The term "meaningful trajectories" means a group of trajectories that lies in a space in which they can be compared. Considering that we are employing a moving sensor, if we directly extract a trajectory of a person on the image, the extracted 2D motion will make no sense. Instead, our approach is to calibrate and compensate the motion of the camera in order to extract world coordinate trajectories that could be even measured in meters.

In Chapter 3 instead we look at the problem of linking target paths across cameras. This problem is addressed in the literature as a problem of matching people across views. So, we assume that we have detected or tracked the target in a camera and that we have a blind-zone in which the target is not visible; then he will reappear in a different camera and must be associated with the previous instance. This problem is called person re-identification by the computer vision community.

Face analysis

With the methods reported in Chapters 3 and 4 we are able to follow groups of people and recover world coordinate trajectories. In this part we exploit the results of these methods to perform face analysis. In Chapter 4 we argue that face detection alone cannot provide a sufficient statistic since each detection is not unambiguously associated with one subject. Considering

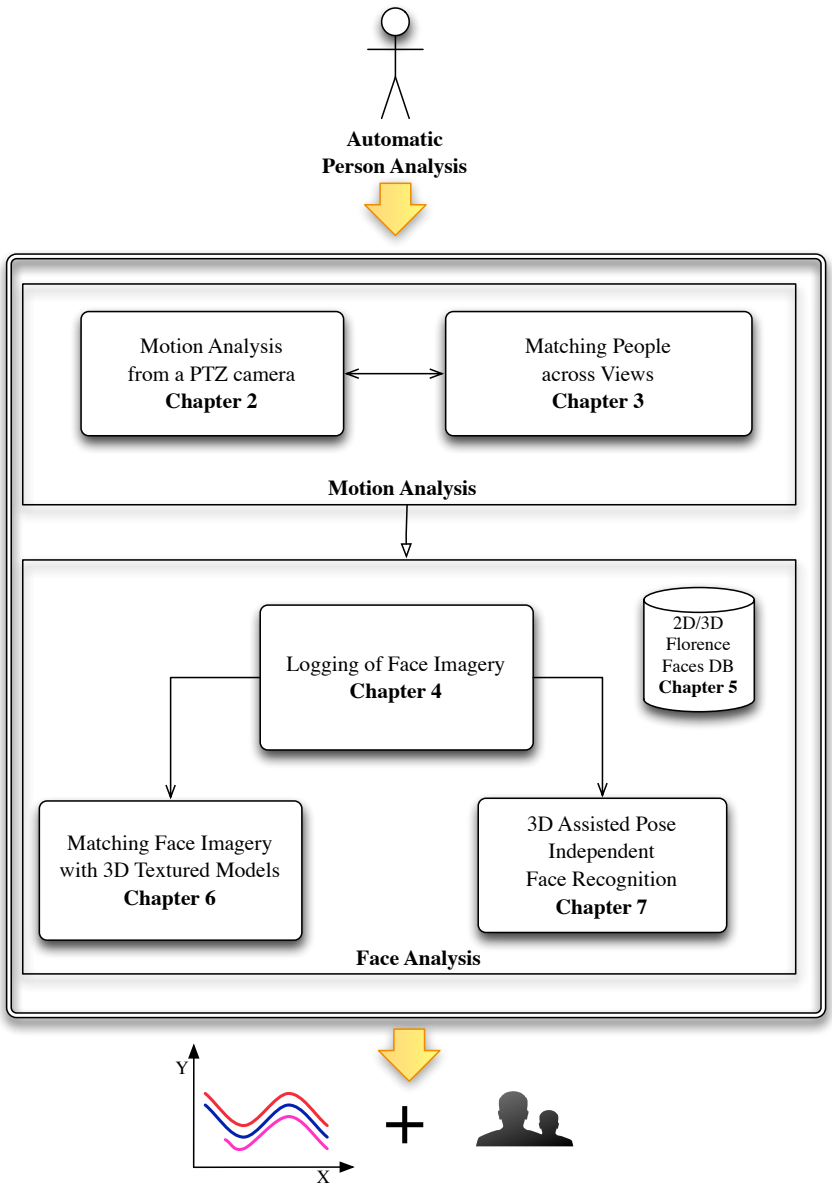


Figure 1.3: Schematic view of the organization of this thesis. From top to bottom, we can see that by observing the subject we analyze the motion and then focus on his face. Each block in the figure corresponds to a chapter of this manuscript. The final output is the extracted motion in term of trajectories and the labels that arise when a face recognition algorithm is applied to multiple faces.

this, in Chapter 4 we propose a novel method to detect frontal and non-frontal faces and to track them in order to log only the most useful face images. We call this procedure face logging and its output is a set of grouped, relevant and high quality face images.

In Chapter 5 we explain the procedure that we use to collect a database of 2D imagery and 3D models, namely the "2D/3D Florence Faces Database", which provides a set of real 3D models that establish a prior that is further used in Chapters 6 and 7 to perform 3D-assisted face recognition.

In Chapter 6 we consider the problem of face recognition in imagery captured in uncooperative environments from a PTZ camera. For each subject enrolled in the gallery, we acquire one high-resolution 3D model from which we generate a series of rendered face images of varying viewpoint. The result of regularly sampling face pose for all subjects is a redundant basis that over represents each target that is used to classify the 2D probe face image.

In Chapter 7 we look at the problem of pose invariant face recognition with a gallery set containing one frontal face image per enrolled subject while the probe set is composed by just a face image undergoing pose variations. Differently from current methods, we solve this problem in a fully automatic way without any manual labeling. In Chapter 7 face recognition is also assisted by 3D information.

Finally, in Chapter 8 we draw some conclusions and provide a summary of the contributions; in Appendix A we summarize some details on other datasets collected during the PhD and in Appendix B we report the publications associated with this work.

1.3 Contributions

The main contributions of this thesis are reported separately for each chapter and are the following:

1.3.1 Motion Analysis

Chapter 2: Motion Analysis from a PTZ Camera

In Chapter 2 we present a novel system that, under reasonable assumptions holding in many real contexts, allows continuous, adaptive, real-time self calibration of a PTZ camera. This enables tracking of multiple targets in wide areas with a single camera, even in the presence of changes of the scene

appearance and for long periods of operation. The approach is thus able to extract world coordinate trajectories of people even if the PTZ camera is moving. The approach provides an adaptive representation of the scene under observation in order to continuously localize itself in the scene. Landmarks are continuously updated in the scene map to account for the changes of the environment due to illumination changes and objects entering/exiting or changing position in the scene. This permits continuous camera calibration over hours of activity as discussed in [45].

From the estimation of camera pose we can compute the relationship between target positions in the 3D world plane and positions in the 2D image and infer the expected imaged height of a target at any image location. This improves both precision/recall performance of human detector algorithms as well as execution time by searching for targets exactly where they are expected to be found.

Finally, since camera motion is compensated, an absolute reference system can be used and tracking can be performed in the world plane, instead of in the image plane. This allows real-time tracking with high and stable accuracy also at large distances and high zoom levels.

Chapter 3: Matching People across Views

The method covered by Chapter 2 is essential for extracting long trajectories but it can not handle with a target that walks around a corner and it cannot link two trajectories between two non-overlapping cameras. To this end, in Chapter 3, we present several approaches solving this problem of matching people across view. We introduce a novel feature descriptor which is shown to be robust for use in person re-identification problems without the additional complexity of feature descriptions that rely on background models or parts. We show how to use an iterative approach to extend sparse discriminative classifiers to rank a large number of candidate individuals. To the best of our knowledge, we are the first to formulate person re-identification as an iterative process of sparse classification, refinement, and ranking. We show through extensive experimental evaluation how our approach is suitable for all re-identification modalities, that it outperforms most approaches at the state-of-the-art, and that it is very competitive with metric learning techniques without requiring labeled training data for learning.

1.3.2 Face Analysis

Chapter 4: Logging of Face Imagery

The main contribution of Chapter 4 is a principled method to associate and group face detection responses from a video stream. The result is a set of clustered, high quality face images. The other contributions are the development of a multi-pose face detector, based on the AdaBoost face detector of Viola and Jones [87,123], that incorporates the outputs of lateral and frontal face detectors to improve robustness to face pose variations and provides a rough estimate of face pose usable as a facial quality measure; the definition of a multi-target tracking and data association module handling multiple intrusions over long periods of time; and the proposal of two distinct facial quality measures, in addition to resolution, which are compared against the brute force approach that logs each detection associated with a track.

The integration of these novel components, along with state-of-the-art components for object detection and tracking, outperforms other face logging methods described in the literature.

Chapter 5: 2D/3D Florence Faces Dataset

The novelty of Chapter 5 is that we are the first to propose, collect and share a dataset that jointly collect 2D face imagery information along with high resolution 3D models. A preview of this dataset is available under request at `http://www.micc.unifi.it/vim/datasets/3d-faces/`[2].

Chapter 6: Matching Face Imagery with 3D Textured Models

The key contribution of Chapter 6 is a new method to perform face recognition between 2D and 3D data. In this way a subject is enrolled with a single-shot procedure that captures a 3D model and thus enrollment becomes very quick. Then, by leveraging the 3D models, the method is able to synthesize poses for the subjects that are used to recognize faces in uncooperative environments using PTZ cameras. While directly collecting images over varying pose for all enrolled subjects is prohibitive at enrollment, the use of high speed, 3D acquisition systems allows our face recognition system to quickly acquire a single model and generate synthetic views offline.

[2]Many thanks to Andrea Ferracani who designed and developed the website.

Chapter 7: 3D Assisted Pose Independent Face Recognition

In Chapter 7 our effort is to improve the performance of standard face recognition algorithms across pose. The novel aspects are the following: we are the first to propose a way to automatically label and semantically align a set of 3D face models without any manual intervention. This solves the landmark ambiguity problem reported previously [7, 134]. We further propose a novel strategy to fit a non-rigid transformation on a face image which exploits the face modeling power of the 3D Morphable Model (3DMM). Similarly to [134], the 3DMM is efficiently fit on a image using a Ridge Regression, that globally preserves the face shape while locally minimizing the landmark reprojection error. By exploiting the fitting, instead of computing LBP on a uniform grid [4], we localize the LBP histograms on the deformed vertices. This gives more precision to the method and obtains feature vectors of the same dimension irrespective of the image size.

Part I

Motion Analysis

Chapter 2

Motion Analysis from a PTZ Camera

Pan-tilt-zoom (PTZ) cameras are powerful to support object identification and recognition in far-field scenes. In this chapter, we present a solution that provides continuous estimation of the camera pose and mapping from scene landmarks, which is robust to rapid camera motion, changes of the environment due to illumination or objects, and scales beyond thousands of landmarks. The estimation of camera pose permits to compute the relationship between the positions of the targets in the 3D world plane and their corresponding positions in the 2D image, and to perform real-time tracking of multiple targets in world coordinates. Since camera motion is compensated, an absolute reference coordinate system can be used and targets can be tracked in the world plane with high and stable degree of accuracy also at large distances and any zooming level. We provide experimental evidence that the solution proposed is capable of supporting effective real-time tracking of multiple targets with PTZ cameras.[1]

[1]This chapter has been published as "Device-Tagged Feature-based Localization and Mapping of Wide Areas with a PTZ Camera" in *in Proc. of CVPR Int'l Workshop on Socially Intelligent Surveillance and Monitoring, San Francisco,USA, 2010.*

2.1 Introduction

Pan-tilt-zoom (PTZ) cameras are powerful to support object identification and recognition in far-field scenes. They are equipped with adjustable optical zoom lenses that can be manually or automatically controlled to permit both wide area coverage and close-up views at high resolution. This capability is particularly useful in surveillance applications to permit tracking of multiple targets in high resolution and zooming in on biometric details of parts of the body in order to resolve ambiguities and understand target behaviors.

However, their practical use in automated video surveillance is complicated by the fact that with this type of camera real time localization and tracking of targets is challenging. On one hand, due to panning, tilting and zooming, the geometrical relationship between the camera view and the 3D observed scene is time-varying. So, in order to provide precise localization and tracking of moving targets, at each time instant this relationship must be estimated and camera motion compensated. This is a complex and time consuming operation. Values from the PTZ camera motor encoders can be exploited but these measurements are too inaccurate to be used effectively [130].

On the other hand, due to the mode of operation we cannot assume any fixed background for target detection. Besides, over long periods of operation, real world scenes also exhibit changes either due to objects or to context variations (Fig. 2.1). According to this, well known detection methods based on background subtraction cannot be used, and some adaptive representation of the scene under observation is also necessary.

Moreover, since we must expect that the imaged targets undergo frequent scale changes, classical target detectors that use specialized classifiers [40, 49, 54] are inadequate, since they would require too much computational effort to permit real-time operation.

In the following, we present a solution that provides continuous adaptive calibration of a PTZ camera and enables effective real-time tracking of multiple targets, so to fully exploit the superior capabilities of PTZ cameras for surveillance applications in large areas. In an offline stage, we collect scene landmarks from a finite number of keyframes taken from different viewpoints. At run time, at each time instant, we estimate the homographic transformation between the camera view and the 3D world plane by matching the keypoints in the current view with those extracted from the keyframes in the offline stage. As a result, camera motion is compensated and the relation-

ship between target positions in the 3D world plane and positions in the 2D image is continuously maintained. Changes in the scene that have occurred over time due to illumination or objects are accounted for with an adaptive representation of the scene under observation that models and updates the uncertainty in landmark localization. Tracking of targets is performed in real world coordinates with a high and stable degree of accuracy, as well as with continuity in the presence of occlusions when targets exit the scene for a few frames.

Figure 2.1: Real world scene with changes due to objects and illumination variations. Planar mosaics from a PTZ camera.

2.2 Related work

In the following, we review research papers that are relevant to the scope of this work and have connections with the method proposed. In particular, we review solutions for self-calibration and target tracking with moving and PTZ cameras.

PTZ camera self-calibration

Self-calibration of moving cameras has been the subject of several research papers. Both offline and online methods have been proposed, some of which have specifically considered PTZ cameras.

A paper by Hartley [67] first demonstrated the possibility of performing offline camera self calibration based on only image content for moving cameras undergoing pure rotation. This method was later extended in many different ways (see [41]) and has also been assumed as the central building block of mosaicing [31, 112]. More recently, Sinha and Pollefeys [113] used the same method to perform offline self-calibration of PTZ devices in a camera network. For each camera, they used feature matching and bundle adjustment to compute the approximate values of the intrinsic parameters of a set of view images taken at different pan and tilt angles and the lowest zooming level. Afterwards, other view images at increasing zooming levels are acquired and processed in the same way to estimate the intrinsic camera parameters for the full range of zooming. The mosaics of the camera views are hence matched each other, and the extrinsic camera parameters are estimated for each camera.

Solutions to online estimation of camera pose for moving and PTZ cameras were proposed respectively in [80, 127] and [38, 90]. In [80], Kleyn and Murray presented a method for real time pose estimation of a moving camera, where they applied online bundle adjustment to the five nearest keyframes sampled every ten frames of the sequence. Unfortunately, this approach cannot be used with PTZ cameras when applied to monitor large areas, since it is likely to produce over-fitting in the estimation of the camera parameters. In [127], re-localization of the camera was performed by using a randomized lists classifier to find the correspondences between the features in the current image and pre-calculated features from all the possible views of the scene, and hence applying RANSAC to obtain the correct camera pose. However, their experiments were performed in indoor environments and the scene under observation was restricted to a relatively small area. Civera et al. [38], proposed a method to perform real-time sequential mosaicing of a scene with a rotating camera. They used Extended Kalman Filter (EKF) and SLAM to estimate the location and orientation of the camera. They only considered the case of camera rotations, and did not account for any zooming operations. The landmarks extracted from the scene were included in the filter state. Due to this, this solution cannot scale with the number of features, and

is likely to show very poor accuracy as the number of features grows beyond a few hundred. Lovegrove et al. [90] claimed to provide full PTZ camera self-calibration but didn't demonstrate self-calibration with variable focal length. As an alternative to using EKF sequential filtering, these authors suggested to use keyframes to achieve scalable performance. The camera parameters between consecutive images were obtained by simple image alignment.

The main limitations of these approaches are that they all assume that the scene is almost stationary that changes are only due to camera motion and there is no zooming operation. These assumptions are clearly inappropriate for tracking with PTZ cameras in real applications.

A few authors have attempted to solve the so-called hand-eye calibration problem [121] using the information provided by the built-in motor actuators of PTZ cameras. The authors of [130] defined a PTZ camera model that explicitly accounts for the drift of camera calibration over many hours of operation by reflecting how focal length and lens distortion vary as a function of zoom scale. A few images were used for initialization, followed by a nonlinear optimization. Results have shown a better calibration accuracy with respect to [113], especially at high zoom scales. However, as reported by the authors, the system fails when a large component in the scene has been moved or when the background is changing slowly. A similar hand-eye calibration strategy was also applied in [114], but accounted for pan and tilt camera movements only.

Tracking with PTZ cameras

General tracking methods cannot perform well on sequences from PTZ cameras. This is mainly due to the fact that with moving sensors targets undergo large and abrupt scale changes that cannot be handled unless the method exploits some prior knowledge or calibration on the scene under observation. Moreover, if camera motion is not compensated, the motion of targets cannot be distinguished from the motion of the camera. Due to this and because of the difficulty of granting at the same time precise online camera self-calibration and real-time tracking, most of the solutions proposed for PTZ camera tracking that have appeared so far were limited to either unrealistic simplified contexts of application or restricted special domains. Several authors assumed scenarios with a single moving target [68, 120]. In [68], Hayman et al. used the affine transform applied to lines and points on a fixed scene background to adapt the PTZ camera focal length so to com-

pensate the changes of size of targets. In [120], Tordoff et al. adjusted the PTZ camera focal length by considering the relative depths of new and old camera poses, estimated according to geometrical constraints.

Other methods were targeted to a specific domain of application [16, 97, 110, 135] or exploited context-specific fiducial markers to obtain an absolute reference and compute the time-varying relationship between the target positions in the 2D image and the positions in the 3D world plane [16, 110, 135]. In [97], hockey players were tracked in a PTZ camera sequence using a detector specialized for hockey players trained with Adaboost and particle filtering based on the detector's confidence. The changes in scale of the targets was managed with simple heuristics using windows slightly larger/smaller than the current target size. This solution was improved in [135] by exploiting motion compensation of the PTZ camera. The authors used the a-priori known circular shape of the hockey rink and playfield lines to locate the reference points needed to estimate the world-to-image homography and compute camera motion compensation. Beyond the fact that these solutions are domain-specific, fiducial markers are likely to be occluded and impair the quality of tracking.

A PTZ camera tracking solution based on motion compensation was proposed in [117] by Kumar et al.. They introduced a layered representation in which spatial and temporal constraints on shape, motion, and layer appearance are modeled and jointly estimated. However, adaptation to the changes of the size of the moving targets was simply modeled by allowing the variance to change according to the target shape. This approach is therefore likely to fail in the presence of abrupt scale changes.

2.2.1 Contributions

In this chapter we present a novel solution that, under reasonable assumptions holding in many real contexts, allows continuous adaptive, real-time self calibration of a PTZ camera, so to permit tracking of multiple targets in wide areas with a single camera, even in the presence of changes of the scene appearance and for long periods of operation. We provide several contributions:

- We develop a Bayes-optimal algorithm for estimating PTZ camera pose (including focal length) and mapping from scene landmarks. This results in a principled approach which is robust to rapid and unpredictable camera motion and scales beyond thousands of landmarks.

- We provide an adaptive representation of the scene under observation. Landmarks are continuously updated in the scene map to account for the changes of the environment due to illumination changes and objects entering/exiting or changing position in the scene. This permits continuous camera calibration over hours of activity as discussed in [45].
- From the estimation of camera pose we can compute the relationship between target positions in the 3D world plane and positions in the 2D image and infer the expected imaged height of a target at any image location. This improves in both precision/recall performance of the detector as well as execution time by searching for targets exactly where they are expected to be found.
- Since camera motion is compensated, an absolute reference system can be used and tracking can be performed in the world plane, instead of the image plane. This allows real-time tracking with high and stable degree of accuracy also at large distances and zooming levels.

Experimental results are presented that validate the method in comparison with other solutions and demonstrate that it improves with respect to the state of the art in tracking with PTZ cameras.

Some of the ideas for calibration and tracking were also used under simplified assumptions in [46]. In this paper, differently from the solution subsequently described, targets were detected manually in the first frame of the sequence and the scene was assumed not to change through time. The method proposed was not robust to rapid camera motion and could not maintain camera calibration over hours of activity.

2.3 Camera Pose Estimation and Mapping

In the following, we introduce the scene model and define the variables used. Then we discuss the offline stage, where a scene map is obtained from the scene landmarks of the keyframes, and the runtime operation, where we perform continuous camera pose estimation and updating of the scene map. Target localization in 3D world coordinates and multiple target tracking are hence discussed in Sect. 2.4.

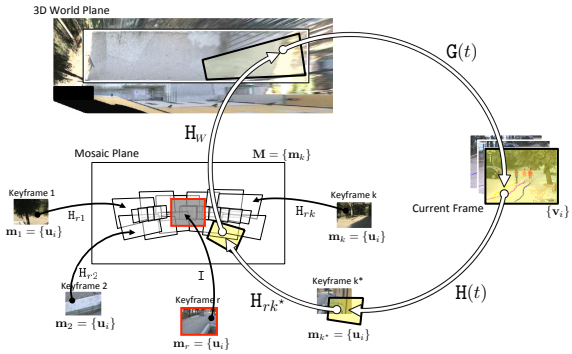


Figure 2.2: Target localization in world coordinates from a PTZ camera, main relationships and elements: the current frame and the landmark observations extracted; the view maps including the scene landmarks; the initial scene map obtained from the union of the view maps; the 3D scene.

2.3.1 Scene model

We consider an operating scenario where a single PTZ camera is allowed to rotate and zoom around its nodal point, observing persons that move over a planar scene. The following entities are defined as random variables:

- The *camera pose* $\mathbf{c}$. Camera pose is time varying, i.e. $\mathbf{c}(t)$. Following [31, 82], we have defined the camera pose only in terms of the (time varying) pan and tilt angles (ψ and ϕ, respectively), and focal length f of the camera. In fact, the principal point is a poorly conditioned parameter, and more precise calibration is obtained if it is assumed to be constant when pan, tilt and focal length are allowed to vary.
- The *scene landmarks* $\mathbf{u}$. These landmarks account for salient points of the scene background and are initially detected in keyframe images in the offline stage using SURF [17]. Keyframes are sampled at fixed intervals of pan, tilt and focal length. The SURF descriptor is maintained associated to each landmark. During online camera operation, since the scene background will modify due to content or illumination variations, new landmarks will be detected while others will be dropped. According to this, we should assume $\mathbf{u}(t) = [x(t), y(t)]$, during runtime camera operation.
- The *view map* $\mathbf{m}$ and *scene map* $\mathbf{M}$. A view map is created for each

keyframe that collects the scene landmarks detected (i.e. $\mathbf{m} = \{\mathbf{u}_i\}$). The scene map is obtained as the union of all the view maps and collects all the scene landmarks that have been detected in the entire scene at different pan, tilt and focal lengths values (i.e. $\mathbf{M} = \{\mathbf{m}_k\}$). Since the scene landmarks change through time, these maps will change accordingly. Due to this, we assume $\mathbf{m}(t)$ and $\mathbf{M}(t)$, during runtime camera operation.

- The *target state* $\mathbf{s}$. The target state is represented in 3D world coordinates and includes both the position and speed of the target. It is assumed that targets move on a planar surface, i.e. $Z = 0$, so that $\mathbf{s} = [X, Y, \dot{X}, \dot{Y}]$.
- The *landmark observations* $\mathbf{v}$. These landmarks account for the salient points that are detected at the current frame. They can either belong to the scene background or to targets. The SURF descriptors of the landmark observations $\mathbf{v} = [x, y]$ are matched with the descriptors associated to the scene landmarks $\mathbf{u}$ registered in the scene map, in order to estimate the camera pose and derive the correct transformation between the current view and the 3D scene.
- The *target observations* in the current frame, $\mathbf{p}$. This is a location in the current frame that is likely to correspond to the location of a target. At each time instant t there is a non-linear function $\mathbf{g}$ relating the position of the target in world coordinates to the location $\mathbf{p} = [x, y]$ of the target in the image. Its estimation depends on the camera pose $\mathbf{c}$ and the scene map $\mathbf{M}$ at time t.

Figure 2.3: Use of bundle adjustment for the estimation of the camera focal length in a sample sequence of a PTZ camera monitoring a large area with right panning and progressive zooming-in. The focal length is estimated at the last frame of the sequence (evidenced with a small square box on the scene mosaic). *Left*: with focal length estimated by online bundle adjustment taking 1 frame every 10 of the sequence (741.174 pixels). *Right*: with focal length estimated by offline bundle adjustment (2097.5 pixels). The true focal length of the last frame of the sequence is 2085 pixels.

Camera localization and mapping requires inference of the joint probability of the camera pose $\mathbf{c}(t)$ and scene landmark locations in the map $\mathbf{M}(t)$, given the landmark observations $\mathbf{v}$ until time t and the initial scene map:

$$p\big(\mathbf{c}(t), \mathbf{M}(t) | \mathbf{v}(0:t), \mathbf{M}(0)\big). \tag{2.1}$$

In order to make the problem scalable with respect to the number of landmarks, Eq. (2.1) is approximated by decoupling camera pose estimation from map updating:

$$\underbrace{p\big(\mathbf{c}(t)|\mathbf{v}(t), \mathbf{M}(t-1)\big)}_{\text{camera pose estimation}} \underbrace{p\big(\mathbf{M}(t)|\mathbf{v}(t), \mathbf{c}(t), \mathbf{M}(t-1)\big)}_{\text{map updating}} \tag{2.2}$$

We use this model to derive a relationship between the target position in the 2D image and its position in the 3D world plane. Fig. 2.2 provides an overview of the relationships between the main entities used to perform runtime target localization in the 3D world plane.

2.3.2 Scene Map Initialization

Scene map initialization is done in an offline stage. We perform a uniform sampling of pan and tilt angles and focal length and take a keyframe at each sample so to have a complete representation of the scene under observation. We also register the coarse values of the pan, tilt and focal length, as provided by the camera actuators. For each keyframe we extract SURF keypoints [17], and create a view map $\mathbf{m}$ that collects all the scene landmarks detected in it, and the camera parameters estimated. The scene map $\mathbf{M}$ is hence obtained as the union of the view maps $\mathbf{m}$.

According to [31], we estimate the optimal values of the external camera parameter matrix $\mathtt{R}_k$ and the internal parameter matrix $\mathtt{K}_k$ at the k-th keyframe, by applying offline bundle adjustment to the sampled keyframes. Differently from [80], where bundle adjustment is performed online applied to a small subset of the frame sequence, this solution exploits the complete scene representation and avoids over-fitting camera parameters, that is a particularly critical phenomenon when the PTZ camera is used in large areas. Evidence of this fact is clearly visible in Fig. 2.3.

Given a reference keyframe and the corresponding view map $\mathbf{m}_r$, the homography that maps each $\mathbf{m}_k$ to $\mathbf{m}_r$ can be estimated as in the usual way for planar mosaicing:

$$\mathtt{H}_{rk} = \mathtt{K}_r \mathtt{R}_r \mathtt{R}_k^{-1} \mathtt{K}_k^{-1} \tag{2.3}$$

2.3.3 Continuous Homography Estimation

When the PTZ camera is in operation, at each time t we use the absolute pan tilt and zoom positional values provided by the camera actuators in order to retrieve the keyframe closest to the current view. These values do not have the required accuracy for the estimation of the camera pose. They are not synchronized with the video stream and non-repeatable controls, small changes in camera pose during operations or lack of stability of the lens system at high zoom typically affect the precision of the measures. Nevertheless, the view map $\mathbf{m}_{k^\star}$ with the closest values of pan, tilt and focal length, is likely to contain most of the scene landmarks that are also visible in the current view.

Following this step, the descriptors of landmark observations $\mathbf{v}$ detected in the current view are matched against the descriptors of scene landmarks $\mathbf{u}$ in $\mathbf{m}_{k^\star}$, according to the distance ratio criterion of [91]. The homography $\mathtt{H}(t)$ from the current view to $\mathbf{m}_{k^\star}(t)$ is hence estimated with RANSAC. The homography $\mathtt{H}_r(t)$ between the current and the reference view is calculated as:

$$\mathtt{H}_r(t) = \mathtt{H}_{rk^\star} \cdot \mathtt{H}(t). \tag{2.4}$$

It includes the information on camera pose with respect to the reference view.

2.3.4 Scene Map Updating

Changes of the visual environment due to illumination or objects entering, leaving or changing position in the scene, will modify the original map as time progresses. They will determine drifting of camera pose estimation and consequently will affect tracking performance. However, while some of these elements will permanently change the scene content, others will only determine temporary changes. In order to have a continuous updated map of the scene content and not account for temporary elements, we introduce a landmark birth-death process, and perform a recursive optimal estimation of landmark locations at each time instant.

Landmark birth-death process

Landmarks that have been discarded as outliers in the computation of $\mathtt{H}(t)$ are included into $\mathbf{m}_{k^\star}$ as new landmarks, according to a likelihood function.

For each outlier, the likelihood is defined as the ratio between the area of the bounding box of the landmarks that were matched and the area of the new bounding box that also includes the outlier. Scene landmarks that have not been matched for a number of frames are assumed to correspond to temporary elements and are therefore removed from the map. The updated map is used to estimate the camera pose at the next time step.

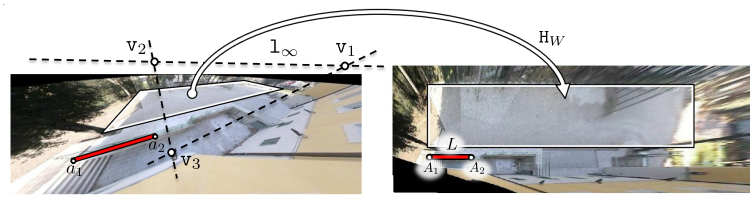


Figure 2.4: The transformation from the 2D mosaic plane (*Left*) to the 3D world plane (*Right*). The vanishing points and the vanishing lines are used for the computation of matrix $\mathtt{H}_p$. A pair of corresponding points in the mosaic and world plane is shown.

Landmark uncertainty modeling and updating

A precise localization of landmarks is obtained by applying the Extended Kalman filtering to the observation model and considering all the possible sources of error that might affect landmark observations.

According to the process described above, at each time t only the view map $\mathbf{m}_{k^\star}$ is updated. Therefore, the map updating factor in Eq. (2.2) can be rewritten as:

$$p\big(\mathbf{m}_{k^\star}(t)|\mathbf{v}(t),\mathbf{c}(t),\mathbf{m}_{k^\star}(t-1)\big) \tag{2.5}$$

By applying Bayes theorem to Eq. (2.5), and assuming that landmark observations $\mathbf{v}$ that match the scene landmarks in $\mathbf{m}_{k^\star}(t)$ are independent of each other, given the scene landmark locations and camera pose, i.e.:

$$p\big(\mathbf{v}(t)|\mathbf{c}(t),\mathbf{m}_{k^\star}(t)\big) = \prod_i p\big(\mathbf{v}_i(t)|\mathbf{c}(t),\mathbf{u}_i(t)\big), \tag{2.6}$$

it results:

$$p\big(\mathbf{m}_{k^\star}(t)|\mathbf{v}(t),\mathbf{c}(t),\mathbf{m}_{k^\star}(t-1)\big) = \prod_i p\big(\mathbf{v}_i(t)|\mathbf{c}(t),\mathbf{u}_i(t)\big)p\big(\mathbf{u}_i(t)|\mathbf{u}_i(t-1)\big) \tag{2.7}$$

where $p\big(\mathbf{u}_i(t)|\mathbf{u}_i(t-1)\big)$ is the prior pdf of the i-th scene landmark at time t given its state at time $t-1$.

Under the assumptions that: scene landmarks $\mathbf{u}_i(t)$ have a Gaussian pdf, the Direct Linear Transform (DLT) is used to compute $\mathtt{H}(t)$ and landmark localizations error have a Gaussian distribution, the observation model $p\big(\mathbf{v}_i(t)|\mathbf{c}(t),\mathbf{u}_i(t)\big)$ corresponds to a linear measurement function of the form:

$$\mathbf{v}_i(t) = \mathbf{H}_i(t)\mathbf{u}_i(t) + \boldsymbol{\lambda}_i(t) \tag{2.8}$$

where $\mathbf{H}_i(t)$ is the 2×2 matrix obtained by linearizing the homography $\mathtt{H}(t)$ at the matched landmark observation $\mathbf{v}_i(t)$ and $\boldsymbol{\lambda}_i(t)$ is an additive Gaussian noise term with covariance $\boldsymbol{\Lambda}_i(t)$.

The covariance of the observation model $\boldsymbol{\Lambda}_i(t)$, can be defined to include all the sources of error that may affect landmark observations, namely: the *landmark transfer error* (arising from landmark spatial distribution and the DLT method), the *landmark uncertainty in the map* and the *keypoint detection error* (originated by the detector). In homogeneous coordinates, $\boldsymbol{\Lambda}_i(t)$ can be expressed as:

$$\mathtt{\Lambda}_i(t) = \mathtt{B}_i(t)\,\mathtt{\Sigma}_i(t)\mathtt{B}_i(t)^\top + \mathtt{H}(t)^{-1}\,\mathtt{P}_i(t)\mathtt{H}(t)^{-\top} + \mathtt{\Lambda}'_i\ , \tag{2.9}$$

where:

- $\mathtt{\Sigma}_i(t)$ is the 9×9 homography covariance matrix that has closed-form expression according to [39] and $\mathtt{B}_i(t)$ is the 3×9 block matrix of landmark observations (in homogeneous coordinates).
- $\mathtt{P}_i(t)$ is the covariance (in homogeneous coordinates) of the estimated landmark position on the nearest view map through $\mathtt{H}(t)$.
- $\mathtt{\Lambda}'_i$ is the keypoint detection error.

The Bayes optimal updating of Eq. (2.7) can be obtained in closed form through multiple applications of the Extended Kalman Filter to each landmark.

The effect of Eq. (2.9) on the estimation of the i-th landmark position is propagated through the Kalman gain, computed as:

$$\mathbf{K}_i(t) = \mathbf{P}_i(t|t-1)\mathbf{H}_i(t)^{-1}\Big[\mathbf{H}_i(t)^{-1}\mathbf{P}_i(t|t-1)\mathbf{H}_i(t)^{-\top} + \boldsymbol{\Lambda}_i(t)\Big]^{-1}. \tag{2.10}$$

2.4 Application to Multiple Target Tracking

In this section, we demonstrate how camera pose estimation can be exploited to perform effective target detection and tracking in 3D world coordinates

with PTZ cameras.

The reference plane of the mosaic (i.e. the image plane of the reference keyframe) is related to the 3D world plane according to a stationary homography:

$$\mathtt{H}_W = \mathtt{H}_s\mathtt{H}_p, \tag{2.11}$$

where $\mathtt{H}_p$ is the rectifying homography obtained by exploiting the single view geometry between the planar mosaic and the scene plane[2], and $\mathtt{H}_s$ is a transformation from pixels in the mosaic plane to 3D world coordinates. The transformation $\mathtt{H}_p$ is obtained from the projections of the vanishing points [86]. The transformation $\mathtt{H}_s$ is estimated from the projection of two points at a known distance L in the world plane onto two points in the mosaic plane (Fig. 2.4).

The function $\mathbf{g}$ mapping the position of a generic target in the world plane onto its position $\mathbf{p}$ in the current frame can be represented through the time varying homography matrix $\mathtt{G}(t)$, in homogeneous coordinates, (see Fig. 2.2):

$$\mathtt{G}(t) = \big(\mathtt{H}_W\mathtt{H}_r(t)\big)^{-1} = \big(\mathtt{H}_s\mathtt{H}_p\mathtt{H}_{rk^\star}\mathtt{H}(t)\big)^{-1}. \tag{2.12}$$

2.4.1 Context-based Target Detection

Camera pose $\mathbf{c}(t)$ and the homography $\mathtt{G}(t)$ calculated at each time instant can be exploited to perform efficient and effective detection of moving targets under reasonable assumptions.

Image Slicing using Geometric Constraints

Assuming that targets remain nearly vertical in the scene, the position $\mathbf{h}$ of the head of the target can be estimated from the feet position $\mathbf{p}$ according to the homology relationship:

$$\mathbf{h} = \mathtt{W}\mathbf{p} \tag{2.13}$$

$\mathtt{W}$ being defined as:

$$\mathtt{W} = \mathtt{I} + (\mu - 1)\frac{\mathbf{v}_\infty \cdot \mathbf{l}_\infty^\top}{\mathbf{v}_\infty^\top \cdot \mathbf{l}_\infty}, \tag{2.14}$$

[2]In the case of a PTZ sensor, the homography between each keyframe and the reference keyframe is the infinite homography $\mathtt{H}_\infty$ that puts in relation vanishing lines and vanishing points between the images.

where $\mathtt{I}$ represents the identity matrix, $\mathbf{l}_\infty$ is the world plane vanishing line, $\mathbf{v}_\infty$ is the vanishing point of the world normal plane direction, and μ is the cross-ratio. The vanishing point $\mathbf{v}_\infty$ is computed as $\mathbf{v}_\infty = \mathtt{K}\mathtt{K}^\top \cdot \mathbf{l}_\infty$, with $\mathbf{l}_\infty = \mathtt{G} \cdot [0,\ 0,\ 1]^\top$ and $\mathtt{K}$ can be derived from $\mathtt{H}(t)$ as in [46] (the dependency on t has been omitted for the sake of clarity in all the expressions above).

Using this information, for each frame, we consider horizontal slices at different vertical positions, such that their height is calculated according to Eq. (2.13) $\pm 10\%$ (Fig. 2.5). The HoG template of [40] is hence applied to each slice appropriately rescaled. Variations of scale of $\pm 10\%$ have insignificant influence on the recall of the detector [122] as shown in Fig. 2.6. This solution allows detection of targets at a single scale with constant rate for each slice, so resulting in computational savings.

Figure 2.5: A sample image (Left) and three horizontal slices (Right). Adjacent image slices have been defined with overlapping regions.

Use of Contextual Information

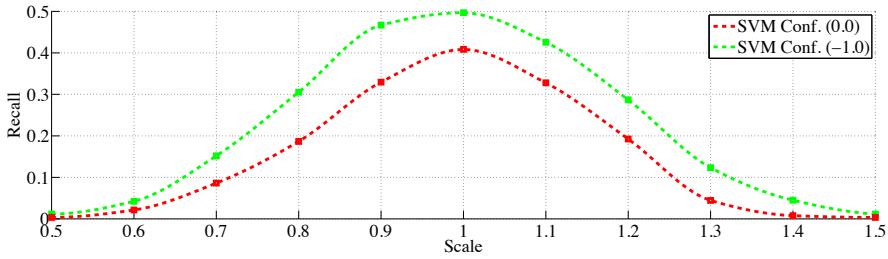


Figure 2.6: Recall performance with image re-scaling using the fixed scale HOG template of [40] on the MICC UNIFI PTZ dataset, with SVM confidence values of 0 and -1. Variations of $\pm 10\%$ in scale don't affect the recall performance substantially.

Since in our case background regions are clearly identified by the presence of matches between scene landmarks and landmark observations, we have exploited this fact and used the probability of false target presence in a window W to discriminate between targets and background regions with human-like patterns:

$$p(\text{target} = \text{false} \mid W) = \int_W \frac{1}{m} \sum_{i=1}^{m} \mathbf{K}\Big(\frac{||\mathbf{w} - \mathbf{v}_i||}{\sigma_{\mathbf{v}_i}}\Big) d\mathbf{w} \tag{2.15}$$

where W is the target detection window, $\mathbf{w} = [x, y] \in W$ are the locations in the detection window, $\mathbf{v}_i$ are the landmark observations matched in W, $\sigma_{\mathbf{v}_i}$ is the scale of $\mathbf{v}_i$, and $\mathbf{K}(\cdot)$ is a Gaussian kernel.

This improves the precision of the detector and reduces the presence of false positives with respect to previous approaches [40] where it is assumed that false positive detections are distributed uniformly in the image [21]. Fig. 2.7 shows an example of detection using Eq. (2.15).

Figure 2.7: Target detections using probabilities of Eq. (2.15): blue regions indicate zones where scene background is highly probable and red regions where targets are more likely to be found.

2.4.2 Multiple Target Tracking in World Coordinates

The relationship of Eq. (2.12) permits target detections in the image plane be in correspondence with their real positions in the world plane at each time instant t. Similarly, the target predictions in the world plane can be put in correspondence with the target observations in the image. Tracking can therefore be performed in an absolute reference coordinate system in

world coordinates with precise separation of the motion of the targets. In order to perform multiple target tracking we follow a two-stage association process using both appearance and motion information.

Greedy Update of Target Track Templates

Active target tracks in the world plane are represented with the color spatiogram of the template of the last target detected. New detections update the track representations according to a greedy, threshold-based approach, using the likelihood function:

$$\gamma_{ij} = \gamma_{ij}^{a} \cdot \gamma_{ij}^{m}, \tag{2.16}$$

where for each target detected i and active track j are considered both the Mahalanobis-Bhattacharyya distance [25] γ_{ij}^{a} between the color spatiograms of the target detected and the active track, and the Mahalanobis distance γ_{ij}^{m} between the position of the target detected and the predicted position for the track calculated in the image plane. As a new detection is associated to an active track, the track template is updated with the template of the detected target. If a new detection is not associated to any active target track, a new track is initiated. If an existing active track has not been linked with any new detections, it is terminated.

Soft Association of Observations to Target Tracks

For each active target track we use the Extended Kalman Filter to estimate the new position of the target in the world plane. The observation model for each target is defined as:

$$\mathbf{p}(t) = \mathbf{g}\big(\mathbf{s}(t)\big) + \zeta(t), \tag{2.17}$$

where $\mathbf{s}$ is the target state, represented in 3D world coordinates, and $\mathbf{g} : \mathbb{R}^4 \mapsto \mathbb{R}^2$ is a measurement function from the world space to the image space defined as:

$$\mathbf{g}\big(\mathbf{s}(t)\big) = \begin{bmatrix} \mathbf{G}(t) & \mathbf{O}_{2\times 2} \end{bmatrix} \mathbf{s}(t), \tag{2.18}$$

with $\mathbf{G}(t)$ being the linearization of Eq. (2.12) and $\mathbf{O}_{2\times 2}$ the 2×2 zero matrix. $\zeta(t)$ is a Gaussian noise term with zero mean and diagonal covariance that models the target localization error in the current frame.

Assuming constant velocity, the motion model in the 3D world plane is defined as:

$$p(\mathbf{s}(t)|\mathbf{s}(t-1)) = \mathcal{N}(\mathbf{s}(t); \mathbf{A}\mathbf{s}(t-1), \mathbf{Q}), \tag{2.19}$$

where $\mathbf{A}$ is the 4×4 constant velocity transition matrix and $\mathbf{Q}$ is the 4×4 process noise matrix. The predicted target positions in the image plane are directly obtained from the target positions predicted in the world plane according to the transformation of Eq. (2.12).

In order to achieve better discrimination between targets that are very close each other, n points are randomly sampled in the neighbourhood of the predicted position and the rectangular template of height given by Eq. (2.13) is extracted at each sample. The three templates with the most similar color spatiograms are used to calculate the probability of association between the k-th observation and the j-th active track. Cheap-JPDAF [55] is used in order to provide an efficient and effective data association:

$$\beta_{kj} = \frac{\gamma_{kj}}{\sum_k \gamma_{kj} + \sum_j \gamma_{kj} - \gamma_{kj} + \kappa}, \tag{2.20}$$

where k is the index of the selected samples, κ is a parameter that models the probability that a target observation is generated by some spurious element and γ_{kj} is computed as in Eq. (2.16). These probabilities are hence used as weights to compute the innovation ν_j of each active track and update the covariance. Innovation is calculated as the weighted combination of the innovations ν_{kj}:

$$\nu_j = \sum_k \beta_{kj} \nu_{kj}. \tag{2.21}$$

Updating of the target covariance $\mathbf{P}_j$ is obtained as:

$$\mathbf{P}_j(t|t) = (1 - \sum_k \beta_{kj})\mathbf{P}_j(t|t-1) + (\sum_k \beta_{kj})\bar{\mathbf{P}}_j(t|t) + \tilde{\mathbf{P}}_j(t) \tag{2.22}$$

where the components respectively account for the uncertainty $\mathbf{P}_j(t|t-1)$ derived from the association between the predicted target position and the sampled location, the uncertainty $\bar{\mathbf{P}}_j(t|t)$ propagated by Kalman Filter state update and the uncertainty $\tilde{\mathbf{P}}_j(t)$ that models erroneous associations.

2.5 Experimental results

The method described above permits continuous real-time self calibration of a PTZ camera and effective real-time tracking in world coordinates of

multiple targets with no fiducial markers. It supports very high and constant precision of target localization even at large distances and any zooming level and operates under assumptions that are verified in most real world contexts. In the following, we present a comparative analysis showing that the state of the art tracking methods do not offer solutions for PTZ cameras with similar characteristics and performance.

A comparative analysis of tracking solutions for PTZ cameras is complicated by several facts. On one hand, the *UBC Hockey* sequence [97] is the only publicly available dataset recorded from a PTZ camera. It is very short and includes frames of a hockey game. So longer and more complex PTZ camera sequences are needed in order to provide meaningful performance assessments. According to this, we have created a new dataset, the MICC UNIFI PTZ dataset [1] including longer PTZ sequences, with several different critical conditions and calibration data associated to each frame. On the other hand, among the few methods that have reported tracking performance figures on the *UBC Hockey* sequence, the method of [97] uses context-specific fiducial markers and the methods in [29] and [132] don't make their code publicly available. So they cannot be assessed on other datasets. Tracking methods that are claimed to have general application can eventually be applied to PTZ camera sequences. Among the most recent and best performing methods we could select only three of them, namely [101], [36] and [133], as their authors were available to a comparative verification.

Considering these facts, we have assessed our method on both the *UBC Hockey* sequence, and the MICC UNIFI PTZ dataset. On the *UBC Hockey* sequence we have compared the performance of our method against the performance reported by [97], [29] and [132]. We also ran the author implementations of [101][3] and [36][4] and reported the performance measured. On the MICC UNIFI PTZ dataset we compared our method against [133], [101] and [36].

2.5.1 Sequences tested

The *UBC Hockey* sequence includes 101 frames of a hockey game. Targets have small size and move erratically and have frequent occlusions. All the targets remain in the scene during the sequence. The scene is observed from a far distance with a large initial tilt angle with respect to the ground plane.

[3] `http://www.ics.uci.edu/~hpirsiav/`

[4] `http://www.eecs.umich.edu/vision/mttproject.html`

The PTZ camera is steered to follow the scene while continuously zooming-in with little pan and tilt.

The MICC UNIFI PTZ dataset contains four PTZ sequences with a total of 3,662 frames and 9,685 annotations of labeled targets. For more details, please see Appendix A.

2.5.2 Characteristics of the methods compared

The authors in [97] use a particle filter to perform tracking in the image plane and a specialized detector trained with Adaboost to detect the hockey players in the sequence.

The methods in [29] and [132] are general tracking methods in the image plane that have reported performance figures on the *UBC Hockey* sequence. Both methods perform target tracking with a particle filter based on detector confidence. The method reported in [29] has been used with the detectors of [40] and [83]. In [132], the authors used Felzenszwalb's part-based detector [54]. Tracking requires learning of a weighting parameter from part of the sequence, and complex hierarchical data association is applied to track multiple targets.

The method by Yang et al. [133] uses the detector of [73] and learns a discriminative part-based appearance model for each target that is continuously updated. Tracking is performed in the image plane by creating tracklets from the association of the detector responses and combining tracklets according to MHT [104].

In the method by Pirsivash et al. [101], detections are performed according to [54]. Target tracking in the image plane is performed by computing the shortest path in a path graph. A greedy algorithm that performs non-maximum suppression of the detector responses is applied to boost tracking performance.

The method by Choi et al. [36] performs multiple target tracking in the world plane with a moving sensor. Target detections are extracted using [54] and target tracking is performed using the MCMC algorithm [75]. The camera pose estimation only accounts for slight variations of the focal length from frame to frame. This method requires that camera pose with respect to the ground plane is available for the first frame. Then camera calibration in the following frames is obtained by sequentially tracking the features with KLT [119]. Camera parameters are hence used to obtain the transformation that relates the target position in the 2D image to its corresponding location

on the ground plane.

2.5.3 Parameter Settings and Metrics

Tests were performed under the following settings:

Camera Pose Estimation and Map Updating Parameters

The interest points were detected and represented using SURF. The RANSAC threshold was set to 3 pixels. The new landmarks observed were tracked for 20 frames and then added to the map. The scene landmark lifetime with no inlier matching was set to 40 frames.

Detection Parameters

Detection is performed using a HoG-based person detector [40]. The detector was forced to optimize recall. A detection is signaled if the confidence is higher than -1.0. New target tracks are initialized for $p(\text{target} = \text{false} \mid W) < 0.05$.

Tracking Parameters

For the outdoor sequences of the MICC-UNIFI PTZ dataset we set the threshold of Mahalanobis-Bhattacharyya distance between the color spatiograms to 0.65 and used a 8-bin spatiogram quantization. For the indoor sequence we used a 0.6 threshold and 16 bin. 100 samples were taken in the neighbourhood of the predicted target position. The lifetime of a track before being terminated was set to 80 frames. Detections are assigned to tracks if the probability of association of Eq. (2.20) is higher than 0.7.

Performance metrics

The accuracy for target detection was evaluated using Recall/FPPI curves and the precision was estimated according to the Multi-Object Detection Precision (MODP) metric, as the average VOC score (the intersection over the union of bounding boxes of ground truth and detection) over all the true positives.

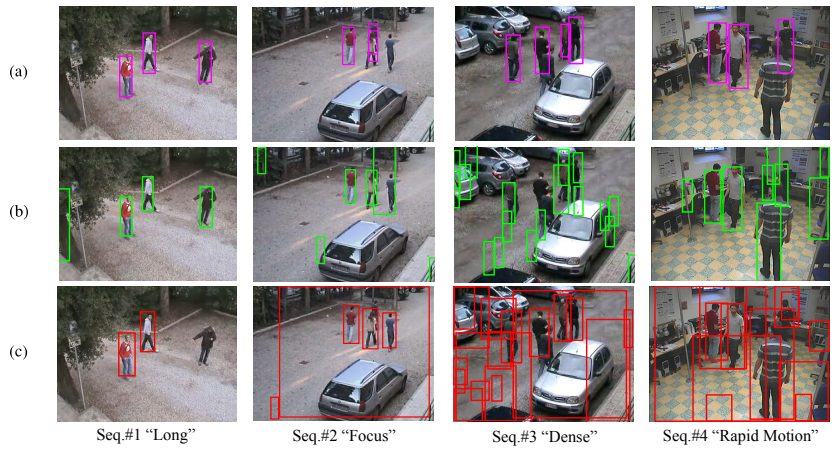


Figure 2.8: Target detections in sample frames of the MICC UNIFI PTZ dataset: (a) our method; (b) Dalal-Triggs' method; (c) Felzenszwalb's method.

The performance for multiple target tracking was evaluated according to the CLEAR MOT metrics [22][5]. The MOTA accuracy index is calculated as:

$$\mathrm{MOTA} = 1 - \frac{\sum_t (\mathrm{TFN}_t + \mathrm{TFP}_t + \mathrm{ID\_SW}_t)}{\sum_t \mathrm{n}_t} \tag{2.23}$$

where TFN_t is the number of tracking false negatives, TFP_t is the number of tracking false positives, $\mathrm{ID\_SW}_t$ is the number of identity switches and n_t represents the true number of targets, in the frame at time t. The precision index MOTP is defined as:

$$\mathrm{MOTP} = \frac{\sum_{i,t} \mathrm{VOC}_{i,t}}{\sum_t \mathrm{TTP}_t} \tag{2.24}$$

where $\mathrm{VOC}_{i,t}$ is the VOC score calculated for the i-th target and TTP_t is the number of tracking true positives, at time t.

[5]We implemented the metrics and we made it publicly available at `https://github.com/glisanti/CLEAR-MOT`.

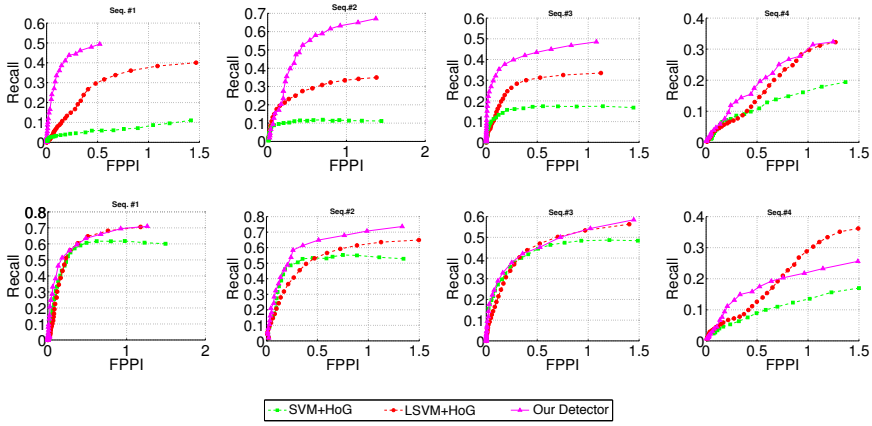


Figure 2.9: Recall/FPPI plots of our method, the Dalal-Triggs' method and the Felzenszwalb's method. Plots obtained from the 3,662 frames of the MICC UNIFI PTZ dataset at low resolution (Top) and high resolution (Bottom).

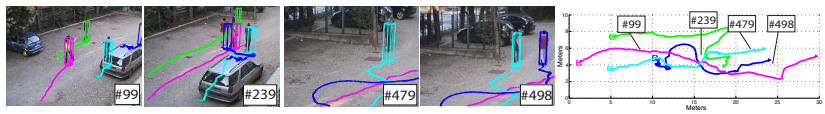
2.5.4 Target Detection

Fig. 2.8 displays examples of target detection on sample frames with our method as expounded in Sect. 2.4.1, in comparison with detections obtained with the direct application of the the Dalal-Triggs' detector [40] (SVM+HoG) and Felzenszwalb's part-based detector [54] (LSVM+HoG).

Fig. 2.9 compares the Recall/FPPI plots evaluated at the original low resolution (top row) and at two-times magnified resolution (bottom row). The average MODP scores are shown in Table 2.1.

It is possible to notice the improvement of detection performance due to the exploitation of camera pose information. The only exception is in Seq.#4 where a target has several partial occlusions. In this case, a better performance is obtained with the LSVM-HoG detector that permits to detect body-parts.

(a) Qualitative results on Seq.#1 "Long".



(b) Qualitative results on Seq.#2 "Focus".

Figure 2.10: Tracking with our method on Seq.#1 (Top) and Seq.#2 (Bottom) MICC UNIFI PTZ dataset. (Left) Sample frames with tracked targets. (Right) Target trajectories.

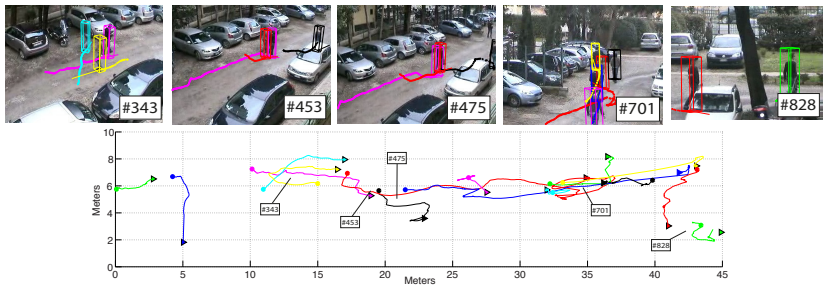


Figure 2.11: Tracking with our method on Seq.#3 MICC UNIFI PTZ dataset (frames from 1 to 900). (Top) Sample frames with tracked targets. (Bottom) Target trajectories. In this sequence, targets move quite randomly with frequent occlusions between targets and make sudden turns and changes of directions.

Seq.	Res.	SVM+HoG	LSVM+HoG	Our Detector
Seq.#1	Low	61.73	63.63	**67.35**
	High	69.09	65.86	**69.19**
Seq.#2	Low	66.89	62.24	**69.18**
	High	64.40	64.04	**65.68**
Seq.#3	Low	68.19	73.87	**74.95**
	High	65.50	74.54	**82.17**
Seq.#4	Low	71.26	76.39	**79.68**
	High	71.85	**75.95**	73.58

Table 2.1: MODP% of people detection methods evaluated over the MICC UNIFI PTZ dataset at low and high resolution for each sequence.

2.5.5 Multiple-Target Tracking

MICC-UNIFI PTZ Dataset

Figs. 2.10, 2.11, 2.12, 2.13 show the target trajectories obtained from our tracking for the four sequences of the MICC UNIFI PTZ dataset. In Table 2.2 we report the CLEAR MOT metrics computed on these sequences in comparison with the methods in [36,101,133]. For a more complete investigation we also report the tracking false positives (TFP%) and false negatives (TFN%) rates, the number of identity switches due to exchange of target identities (ID_SW) and the number of trajectory fragmentations (TR_FR) (typically when a target exits the camera field of view for several frames either due to camera zooming in or random movements of a target with respect to the camera).

The experiments show that our method has the highest accuracy than the other methods with all the sequences. It scores very good accuracy (MOTA) on Seq.#1 and Seq.#2 and has good accuracy also in the more complex scenes of Seq.#3 and Seq.#4.

Our method has the highest precision (MOTP) on Seq.#3 and almost the same precision as [101] on Seq.#1 and Seq.#2. In Seq.#4 The higher precisions of [101], and [36] should be ascribed to the use of the Felzenszwalb's detector [54]. In fact, in this sequence, the camera movements determine frequent partial occlusions of the targets that cause misses with the full body HoG detection of our method.

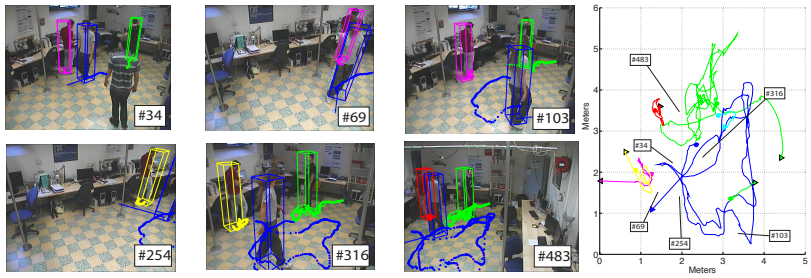


Figure 2.12: Tracking with our method on Seq.#4 MICC UNIFI PTZ dataset. (Left) sample frames with tracked targets. (Right) Target trajectories. In this sequence, targets move quite randomly with frequent occlusions between targets.

Our method scores the lowest trajectory fragmentation score, showing that it can support continuous tracking much better than the others. Tracking in world coordinates permits recovery from critical cases such as when a target exits the field of view for a few frames with no identity switches (see f.e. frames #479 and #498 of Seq.#2 in Fig. 2.10(b)). Trajectory fragmentations occur only in those cases where targets exit the field of view for long periods as in Seq.#3. In these cases they originate a new track when they reenter in the scene.

With our method, tracking false negatives occur when targets have small size (see f.e. frame #475 in Seq.#3 in Fig. 2.11) or are very close to the camera (see f.e. frame #34 of Fig. 2.12). The TFN score is much lower than the methods by [36, 101] in all the sequences. Lower TFN rates of [133] in Seq.#2 and in Seq.#3 are counterbalanced by very high TFP rates in the same sequences. Tracking false positives are likely to occur when the tracker drifts on the background or when a target exits the field of view for a few frames (see frame #254 of Seq.#4 in Fig. 2.12 f.e.). The methods of [36, 101] have a lower TFP rate only apparently. Indeed in these methods, differently from our case, when the detector misses the target for several frames, the tracker doesn't try to recover the track, but terminates the current track and initiate a new one. As a result, an unreasonably high number of trajectory fragmentations is generated.

All the methods show a very low number of identity switches.

	MOTA%	MOTP%	TFN%	TFP%	ID_SW	TR_FR
Seq.#1						
Our approach	**92.0**	66.2	**5.0**	4.4	**0**	**0**
Pirsiavash [101]	84.7	**66.7**	13.42	**0.92**	0	21
Choi [36]	56.0	63.4	41.1	0.94	1	46
Yang [133]	66.7	58.2	7.4	25.2	0	18
Seq.#2						
Our approach	**77.3**	66.8	15.0	6.5	**0**	**0**
Pirsiavash [101]	74.8	**67.3**	22.1	0.85	0	33
Choi [36]	50.5	60	47.1	**0.69**	1	31
Yang [133]	74.7	54.4	**3.5**	20.7	2	18
Seq.#3						
Our approach	**67.5**	**67.8**	28.3	3.5	**1**	**19**
Pirsiavash [101]	63.3	66.8	34.6	**0.42**	0	55
Choi [36]	21.6	65.4	75.6	1.34	2	51
Yang [133]	51.0	65.6	**4.52**	41.4	2	122
Seq.#4						
Our approach	**63.4**	67.3	**26.8**	9.5	**1**	**4**
Pirsiavash [101]	51.1	**74.1**	48.2	**0.63**	0	20
Choi [36]	48.2	71.3	48.1	1.82	0	32
Yang [133]	45.6	66.7	50.17	2.85	1	24

Table 2.2: Tracking performance on MICC UNIFI PTZ dataset.

UBC Hockey Sequence

The performance figures obtained with the *UBC Hockey* sequence are reported in Table 2.3, in comparison with the other methods.

In our method we used the detector of Okuma et al. [97] to detect targets. The camera calibration data, needed to build the scene map $\mathbf{M}(0)$, were taken directly from the sequence, sampling a keyframe every ten frames. It is apparent that our method largely outperforms the methods of [29] and [97] in both precision and accuracy. A lower precision is instead observed with respect to [132]. This is determined by a special unique situation (see frame #38 of Fig. 2.13), where a target track remains stuck to the scene background for five frames, so generating several false positives. The method of [36] did not provide significant results mainly because that method requires that the scene horizon is in the initial image of the sequence, which is a condition that is not satisfied in the *UBC Hockey* sequence.

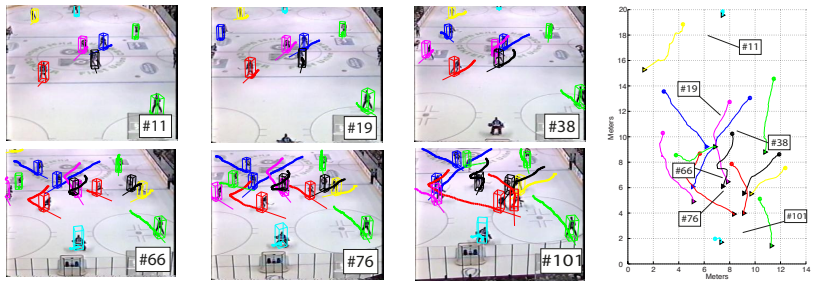


Figure 2.13: Tracking with our method on the *UBC Hockey* sequence (Left): sample frames with tracked targets. (Right) Target trajectories.

	MOTA%	MOTP%	TFN%	TFP%	ID_SW
HOCKEY					
Our approach	91.6	61.3	6.3	2.0	1
Breitenstein [29]	76.5	57.0	22.3	1.2	**0**
Okuma [97]	67.8	51.0	31.3	**0.0**	11
Yan [132]	**91.7**	71.6	**1.76**	6.49	0
Pirsiavash [101]	16.4	**74.4**	82.9	0.19	5
Choi [36]	Tracks targets occasionally				
Yang [133] *	–	–	–	–	–

Table 2.3: Tracking performance on UBC Hockey dataset. *Code not made available by authors.

Operational Constraints and Computational requirements

The solutions in [101], [133] and [132], although having good performance figures in terms of accuracy and precision, nevertheless all require that target detections be available beforehand. So they cannot be employed for real-time tracking with PTZ cameras in real applications. In their experiments on the *UBC Hockey* sequence, the authors of [29] have reported a variable performance between 0.4 and 2 fps on a Dual-core@2.13GHz. On the same dataset, the method in [97] has reported 1fps on a Dual-core@2.66 GHz. The authors of [36] have reported that their method is capable to operate at 5 fps, not accounting for detection. Therefore, we can presume that the total rate of the complete system is less than 2 fps, that is too slow for effective tracking.

Our method has been verified to perform real-time multiple target tracking in sequences from PTZ cameras at 12 fps on Intel Xeon™ Dual Quad-Core at 2.8GHz and 4GB of memory, with no GPU processing. The current implementation of the method exploits multiple cores and was developed in C/CC++. Frame grabbing, camera calibration and context analysis are calculated in one thread and sent to the other threads where detection and tracking are performed. Extraction of target measurements was implemented using the Intel© Threading Building Block library.

2.6 Conclusions

In this chapter, we have presented an effective solution for real-time multiple target tracking from a single PTZ camera observing a planar scene. The solution integrates a complex unified framework for on-line camera calibration, context-based target detection, tracking in world coordinates and multi-stage data association. It maintains a continuous relationship between the target observations in the image plane and the corresponding positions in the world plane, as estimated with online continuous calibration of the camera during operation. This permits improvement of the performance of the detector and allows more effective tracking of multiple targets in world coordinates with camera motion compensation. As a result, the solution proposed allows tracking of multiple targets with PTZ cameras in real-time with high and stable degree of accuracy, also at large distances and any zooming level. It achieves the state of the art performance of tracking with these cameras.

Chapter 3

Matching People across Views

In this chapter we introduce a method for person re-identification based on discriminative, sparse basis expansions of targets in terms of a labeled gallery of known individuals. We propose an iterative extension to sparse discriminative classifiers capable of ranking many candidate targets. The approach makes use of soft and hard re-weighting to re-distribute energy among the most relevant contributing elements and to ensure that the best candidates are ranked at each iteration. Our approach also leverages a novel visual descriptor which we show to be discriminative while remaining robust to pose and illumination variations. An extensive comparative evaluation is given demonstrating that our approach achieves state-of-the-art performance on single- and multi-shot person re-identification scenarios on the VIPeR, i-LIDS, ETHZ, and CAVIAR4REID datasets. The approach is additionally quite efficient, capable of single-shot person re-identification over galleries containing hundreds of individuals at about 30 re-identifications per second.[1]

[1]This chapter has been submitted to the journal *IEEE Transaction on Pattern Analysis and Machine Intelligence* as "Person Re-identification by Iterative Re-weighted Sparse Ranking" and is currently under major revision.

3.1 Introduction

Person re-identification is the task of recognizing a person, captured by one or more cameras, over a range of candidate targets represented as a gallery of already-labeled subjects. This gallery may contain imagery of known subjects from one or more sensors, and there may be no guarantee that an unknown subject has already been imaged from the same point of view as the images to be re-identified. In fact, some of the main issues in person re-identification are due to the fact that the same person is usually acquired at different times, by different disjoint cameras, and this can result in large variations in target appearance because of different illumination conditions, different poses or partial occlusions.

Considering the large number of cameras in typical modern surveillance networks, person re-identification is critical as it is a way of maintaining identity information about targets in multiple views over potentially long periods of time. This matching across cameras is traditionally cast as a retrieval problem, and its formulation as such has a strong impact on how re-identification problems are defined and how performance is evaluated: given one or more images of an unknown target, the re-identification task is to *rank* all individuals in a gallery of known target images in terms of similarity to the person to be recognized.

Much of the research on person re-identification has concentrated on human appearance modeling. A number of descriptors of image content have been proposed to discriminate identities while compensating for appearance variability due to changes in pose, illumination and camera viewpoint. Supervised metric learning methods have also been applied to ameliorate the problems of multi-view appearance changes. Metric learning approaches require labeled training data and most of them also require new training data when camera settings change. Discriminative models, in contrast to appearance-based ones, can suffer from lack of training data in small gallery image sets and are often unsuitable for ordering large numbers of candidates due to their inability to reliably rank all but a few of the best ones.

The literature on person re-identification focuses on several different modalities or scenarios that are recognized as de facto standards for performance evaluation of re-identification algorithms [18,19,52]. These modalities are characterized in terms of how many images of each individual are known *a priori* to be in the gallery and probe sets, and according to whether or not it is known that multiple images in the probe set correspond to a single

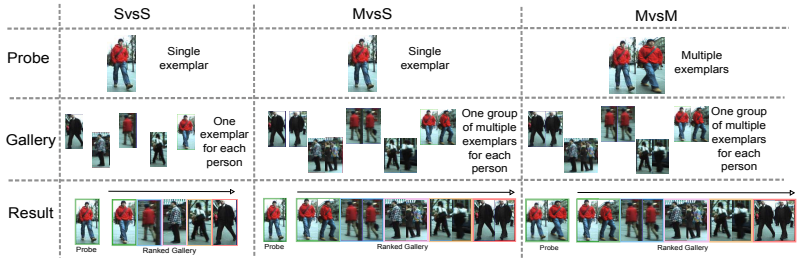


Figure 3.1: Re-identification modalities. Each column represents a modality and each row illustrates the probe, gallery and result sets for each modality. For clarity we illustrate MvsS and MvsM modalities only for the $N = 2$ case. Colored frames in the last row are used to represent groups of images of the same person in the query and result sets.

target. The three most common are: the single-versus-single (SvsS) modality in which a single probe exemplar is matched against individual gallery exemplars; the multi-versus-single (MvsS) modality in which a single probe exemplar is matched against a gallery of groups of multiple exemplars; and the multi-versus-multi (MvsM) modality in which a group of probe exemplars is matched against a gallery of grouped exemplars. These three experimental modalities are illustrated in Fig. 3.1.

In this article we propose a robust and efficient approach to person re-identification that is applicable to all the modalities described in Fig. 3.1. Building on ideas developed for sparse discriminative face recognition [128], our approach uses ℓ_1-regularized sparse basis expansions of queries in terms of a basis of gallery images. Through analysis of reconstruction error we partially rank the gallery in terms of similarity to the query probe, and then re-weight this initial solution in order to mute the response of vectors contributing little to the initial expansion. Through the use of a novel, iterative re-weighting algorithm, we then proceed to rank the remaining gallery individuals through analysis of re-weighted sparse basis expansions. By using sparse reconstruction, rather than learning discriminative models, we are able to preserve multiple aspects of each gallery individual and leverage this information at re-identification time. This also has the advantage that new gallery data can be effortlessly added with no re-training.

The contributions of this work are the following:

- We introduce a novel feature descriptor which is shown to be robust for

use in person re-identification problems without the additional complexity of feature descriptions that rely on background models or parts.
- We show how to use an iterative approach to extend sparse discriminative classifiers to rank a large number of candidate individuals. To the best of our knowledge, we are the first to formulate person re-identification as an iterative process of sparse classification, refinement, and ranking.
- We show through extensive experimental evaluation how our approach is suitable for all re-identification modalities, that it outperforms most approaches at the state-of-the-art, and that it is very competitive with metric learning techniques without requiring labeled training data for learning.

In the next section we review the literature related to person re-identification. Our approach to describing the visual appearance of persons is given in Sec. 3.3, and in Sec. 3.4 we show how to perform re-identification with sparse basis expansions. In Sec. 3.5 we describe our approach to person re-identification using sparse iterative ranking. In Sec. 7.7 we give an extensive comparative evaluation of our technique with respect to the state-of-the-art on four publicly available datasets used in the literature. Finally, in Sec. 4.4 we draw some conclusions and discuss new directions for research.

3.2 Related work

Many recent works have addressed the problem of person re-identification. The majority of techniques are principally divided into *appearance based* and *learning based* approaches.

Appearance-based re-identification: The first work that considered the problem of appearance models for person recognition, reacquisition and tracking was that of Gray et al. [61]. The authors argue that until then these problems had been evaluated independently and that there is a need for metrics that apply to complete systems [59, 60]. They proposed a standard protocol to compare results[2] and introduced the VIPeR dataset for re-identification. The first paper to follow Gray's guidelines was [62]. In this work the authors propose an algorithm that learns a domain-specific similarity function using an ensemble of local features and AdaBoost. Features are raw color chan-

[2]They proposed the Cumulative Matching Curve (CMC) on a standardized dataset and showed that this one curve can be converted to a Synthetic Disambiguation Rate (SDR) for a single camera.

nels in many color spaces and texture information captured by Schmid and Gabor filters.

Descriptors of visual appearance for person recognition can be highly susceptible to background clutter, and many approaches to person re-identification use background modeling [18, 19, 52] or part-based person appearance models [34, 52] to separate foreground from background signals. In [52] the authors use a sophisticated appearance model, the Symmetry-Driven Accumulation of Local Features (SDALF) descriptor that models human body parts by estimating the axis of symmetry of a person and obtaining the head, torso, and legs positions. Each part is then represented by weighted HSV color histograms, maximally stable color region descriptors [56], and recurrent highly-structured patches. This work also applies a strong, generative background prior that enhances the discriminative power of the descriptor by segmenting the person from the background [76]. In [18] and [19] a multi-shot appearance-based method similar to [52] is proposed in order to condense a set of frames of the same individual into a highly informative signature, which they call the Histogram Plus Epitome (HPE). In [34] the authors employ an estimate of body pose to guide the process of feature extraction. They extend the pictorial structure model [53] with their Custom Pictorial Structure (CPS), which is a two-step iterative process that alternates between estimating pose and updating the appearance model.

Another state-of-the-art approach with performance similar to [34] is proposed in [14]. The authors use an appearance model that, in contrast with [52] and [34], does not rely on body parts. The approach is based on a descriptor called the Mean Riemannian Covariance Grid (MRCG) [14], which is an extension of Spatial Covariance Regions (SCR) [15], that is the covariance of a vector of eleven cues derived from equalized RGB colors. The MRCG descriptor is computed as a mean of gallery examples and is only applicable to multi-shot re-identification modalities. The person re-identification problem was extended to groups in [140]. The authors show that groups represent a contextual cue that can be exploited to improve person re-identification.

A problem with many appearance-based methods is that they must compute average representatives of individuals over multi-shot gallery sets. Our approach, on the other hand, is able to exploit multiple gallery examples and in contrast to many appearance-based approaches does not require sophisticated background modeling. Re-identification problems are often charac-

terized by poor and variable image quality on which it can be hard to fit background or part-based models without relying on scene-specific information.

Learning-based re-identification: Re-identification can also be cast as as a learning problem in which either metrics or discriminative models are learned. The authors of [109] propose a discriminative model created using Partial Least Squares (PLS) which weights features according to their discriminative power for each different gallery instance. In [48], a metric learning framework is used to obtain a robust Mahalanobis metric for large margin nearest neighbor classification with rejection (LMNN-R). The approach in [103] is a supervised technique that uses pairs of similar and dissimilar images and a relaxed RankSVM algorithm to rank probe images. Another metric learning approach is that of [81] in which learns a Mahalanobis distance from equivalence constraints derived from target labels.

The Probabilistic Distance Comparison (PRDC) approach [139] introduces a novel comparison model which aims to maximize the probability of a pair of correctly matched images having a smaller distance than that of an incorrectly matched pair. The same authors in [125] model person re-identification as a transfer ranking problem where the goal is to transfer similarity observations from a small gallery to a larger, unlabeled probe set. A set-based discriminative ranking approach was also recently proposed which alternates between optimizing a set-to-set geometric distance and a feature space projection, resulting in a discriminative set-distance-based model [129]. Camera transfer approaches have also been proposed that use images of the same person captured from different cameras to learn metrics [9, 72]. In [13] the authors apply learning in a covariance metric space using an entropy-driven criterion to select the most descriptive features for a specific class of objects. Recently saliency has been considered when matching people across views and a novel method eSDC [107] has been proposed in order to learn saliency parts of a human in a unsupervised fashion.

Learning-based approaches have recently reported higher re-identification accuracy with respect to the state-of-the-art, but usually at the cost of setting aside a portion of available labeled data for learning metrics or requiring large gallery sets for fitting discriminative models. Our approach outperforms the state-of-the-art at rank-1 in all modalities without resorting to learning metrics or fitting discriminative models to gallery image sets. Re-identification problems are often also characterized by a lack of reliably labeled data, and

the need to label image data for each scenario, camera configuration and parameter settings is a disadvantage of metric learning approaches.

Hybrid classification approaches: Though not applied specifically to ranking problems like re-identification, classification methods based on regularized, sparse basis expansion of test samples can be considered hybrid methods that exploit appearance and discriminative modeling at recognition time. The appearance of each gallery exemplar is maintained in a basis, and the residual error of sparse reconstructions is used for discrimination. Wright et al. [128] show how this can be a powerful tool for face recognition. This approach has been extended several times, integrating correntropy [69] and kernel-based sparse reconstruction [78].

In this chapter we propose a way to extend sparse discriminative classifiers to person re-identification problems. Such approaches do not directly generalize to ranking problems like re-identification because ℓ_1-regularized basis expansions, by there very nature, can only support ranking of a small number of individuals. We show how regularized basis expansions can be iterated in order to improve the quality of re-identification and to rank a large number of samples.

3.3 A descriptor for re-identification

We employ a feature representation that requires no foreground/background segmentation and does not rely on body-part localization. The descriptor is designed to be discriminative and at the same time very efficient.

Given an input image of a target (that is a rectangular sub-image of an image containing the target), it is scaled to a canonical size $W \times H$ (64×128 pixels in all our experiments). A spatial pyramid is built by dividing the person image into overlapping horizontal stripes of 16 pixels in height. From each stripe we extract a Hue-Saturation (HS) histogram and an RGB histogram. The use of horizontal stripes allows us to capture information about vertical color distribution in the image, while overlapping stripes allow us to maintain color correlation information between adjacent stripes in the final descriptor. The use of HS histograms results in a portion of the descriptor that is invariant to illumination variations, while the RGB histogram from each stripe captures more discriminative color information especially for targets whose visual appearance tends toward the black or white regions of the color cube. We equalize all RGB color channels before

extracting the histogram.

To eliminate background information we use a straightforward approach that weights a pixel's contribution to its corresponding histogram bins according to an Epanechnikov kernel centered on the target image:

$$K(x,y) = \begin{cases} \frac{3}{4}(1 - (\frac{x}{W})^2 - (\frac{y}{H})^2) & \text{if} \quad |(\frac{x}{W})^2 + (\frac{y}{H})^2| \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

where W and H are the image width and height, respectively, and are the only parameters of the Epanechnikov kernel. This discards (or diminishes the influence of) background information and avoids the need to learn a background model for each scenario. This renders our method more general and more efficient, while maintaining simplicity with respect to techniques that use complex background or part-based models.

Color is only one aspect of target appearance. To the HS and RGB histograms we concatenate a set of Histogram of Oriented Gradients (HOG) descriptors computed on a grid over the image as described in [40]. The HOG descriptor captures local structure and texture in the image that are not captured by the color histograms. In order to extract only gross details of target appearance (e.g. vertical, horizontal and diagonal patterns) we only use 4-bin gradient orientation histograms for each HOG block.

As the final stage of descriptor computation, we take the square root of all descriptor bins. This is a well-known technique in the image classification community [37] and helps reduce the "burstiness" phenomenon of features by discounting the effect of small changes in bins that already have significant weight in them. In preliminary experiments we found this to improve robustness of Euclidean distances between descriptors.

HS histograms contain 8×8 bins, while RGB is quantized to $4 \times 4 \times 4$ bins. Both the HS and RGB histograms are computed for the 15 levels of the pyramid (8 stripes for the first level plus 7 for the second level of overlapping stripes). The result is a total of $1,920$ color histogram bins. To this we concatenate a HOG descriptor in which each block consists of a grid of 2×2 cells of 8×8 pixels. We extract the HOG descriptors from a sub-image obtained by removing 8 pixels from top, bottom, left and right of the original in order to remove background details. For each 8×8 cell we compute the gradient histogram over only 4 angular bins as described above (instead of the usual 9 angular bins). Given the canonical size of 64×128 pixels, the dimension of the HOG descriptor is $1,040$ bins, and the final

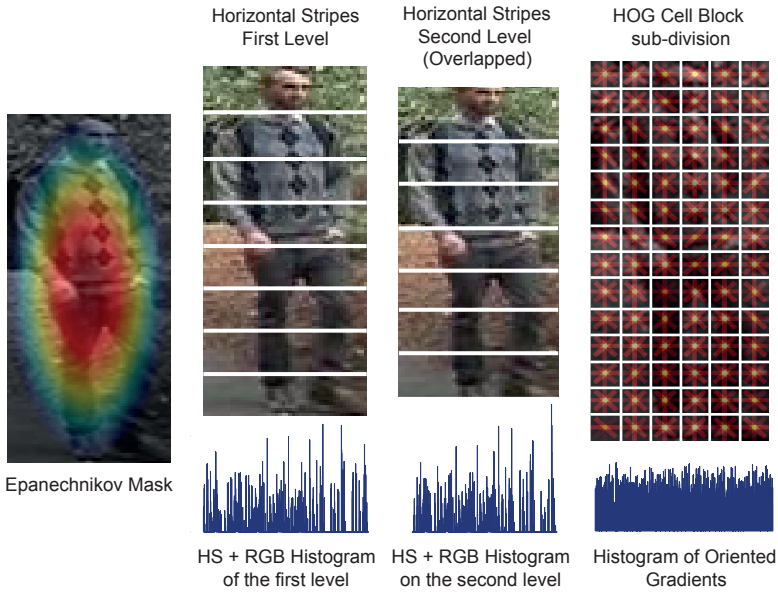


Figure 3.2: The feature extraction process. (a) An Epanechnikov kernel is used to weight the contribution of each pixel to the HS and RGB histograms computed on overlapping stripes (b) and (c). These weighted histograms are concatenated with a grid of overlapping HOG descriptors (d).

descriptor dimensionality of each person image is thus $2,960$. The entire descriptor construction process is illustrated in Fig. 3.2 and an extensive evaluation of descriptor performance is given in Sec 3.6.1.

3.4 Sparse discriminative basis expansions

In this section we describe an approach to using basis expansions for classification and show how this basic approach does not generalize in a straightforward way to problems like re-identification due to its inability to rank all but a few confidently classified individuals. In Sec. 3.5 we introduce an iterative algorithm for ranking with sparse basis expansions that addresses these shortcomings.

3.4.1 Sparse basis expansions for classification

The main idea behind the use of basis expansions for building discriminative classifiers is that, given sufficient training samples $\mathbf{t}_{i,1}, \ldots, \mathbf{t}_{i,n_i}$ from some class i, a test sample $\mathbf{y}$ of the same class should approximately lie in the linear span of the training samples:

$$\begin{aligned} \mathbf{y} &\approx \alpha_{i,1}\mathbf{t}_{i,1} + \alpha_{i,2}\mathbf{t}_{i,2} + \ldots + \alpha_{i,n_i}, \mathbf{t}_{i,n_i} && (3.2) \\ &= \sum_{j=1}^{n_i} \alpha_{i,j}\mathbf{t}_{i,j} && (3.3) \\ &= \mathbf{T}_i\boldsymbol{\alpha}_i && (3.4) \end{aligned}$$

for some optimal choice of scalar coefficients of reconstruction $\alpha_{i,j}$, for $j = 1, \ldots, n_i$. We use $\mathbf{T}_i$ to represent the matrix of basis vectors for class i, and $\boldsymbol{\alpha}_i = [\alpha_{i,1}, \ldots, \alpha_{i,n_i}]^T$ to represent the vector of reconstruction coefficients for the same class.

The general, multi-class basis expansion for C individuals in the gallery then becomes:

$$\begin{aligned} \mathbf{y} &\approx [\mathbf{T}_1\ \mathbf{T}_2\ \cdots\ \mathbf{T}_C]\left[\boldsymbol{\alpha}_1^T\ \boldsymbol{\alpha}_2^T\ \cdots\ \boldsymbol{\alpha}_C^T\right]^T \\ &= \mathbf{T}\boldsymbol{\alpha}. && (3.5) \end{aligned}$$

The basis $\mathbf{T}$ can be highly overcomplete, but if $\mathbf{y}$ is an instance of a person in the gallery we desire that the energy in the basis expansion be concentrated in the relatively few coefficients corresponding to the probe examples corresponding to $\mathbf{y}$. We can impose this sparsity constraint on the solution by formulating it as an ℓ_1-regularized least squares problem:

$$\widehat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \|\mathbf{y} - \mathbf{T}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1, \tag{3.6}$$

where λ controls the tradeoff between minimization of the ℓ_2 reconstruction error and the ℓ_1 norm of the coefficients used to reconstruct $\mathbf{y}$. This formulation is known as Lasso Regression in the statistics literature [118].

Regularized basis expansions of this type are generally referred to as *sparse* because the ℓ_1 regularization term, depending on the sparseness factor λ, tends to cause the coefficients of reconstruction to collapse to zero except for a few important basis vectors. The form of Eq. (3.6) is particularly convenient because it represents a whole class of solutions to the approximate reconstruction problem of Eq. (3.5). When $\lambda = 0$, Eq. (3.6) results in a

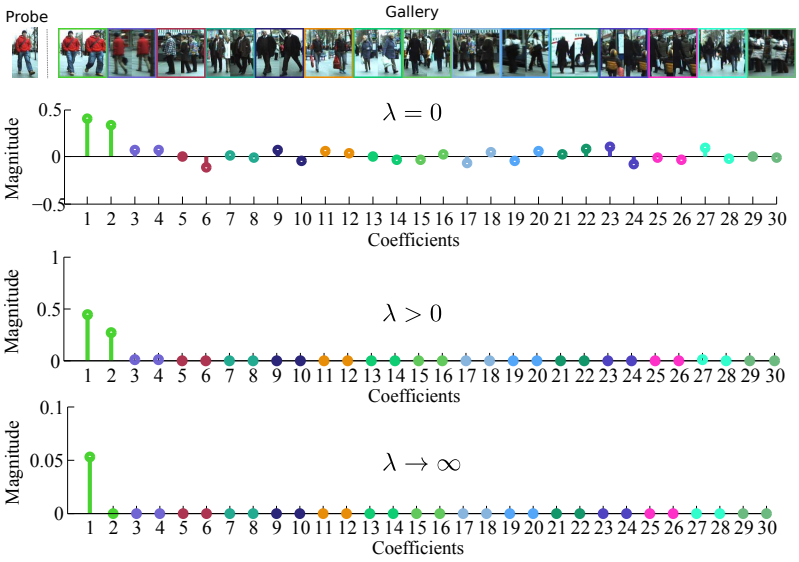


Figure 3.3: Basis-expansion results for MvsS ($N = 2$) re-identification on ETHZ1. Top: (left) probe sample, (right) the first 15 samples in the gallery. Each color represents a single subject which has two instances. Bottom: reconstruction coefficients for least squares ($\lambda = 0$), sparse ($\lambda = 0.2$) and nearest neighbour ($\lambda = 0.6$).

standard least squares solution. For $\lambda > 0$, we obtain solutions of increasing sparseness with increasing λ. Eventually, as $\lambda \to \infty$, only a single non-zero coefficient will be admitted in the solution of Eq. (3.5). We refer to this last solution, with $\lambda \to \infty$, as the *nearest neighbor* solution since only the ℓ_2-closest training sample to $\mathbf{y}$ will have a corresponding non-zero coefficient in $\widehat{\boldsymbol{\alpha}}$. In Fig. 3.3 we illustrate these three types of solutions for a MvsS re-identification problem. The top row of Fig. 3.3 illustrates the probe and gallery images for a re-identification query. The plot in the second row shows the coefficients of a least squares solution ($\lambda = 0$), followed by a sparse solution ($\lambda = 0.2$), and finally the nearest neighbor solution ($\lambda = 0.6$ for this example).

We can derive a decision rule for classification by analyzing the reconstruction error for solutions to Eq. (3.5) restricted to basis vectors corresponding to individual gallery subjects. The normalized reconstruction error

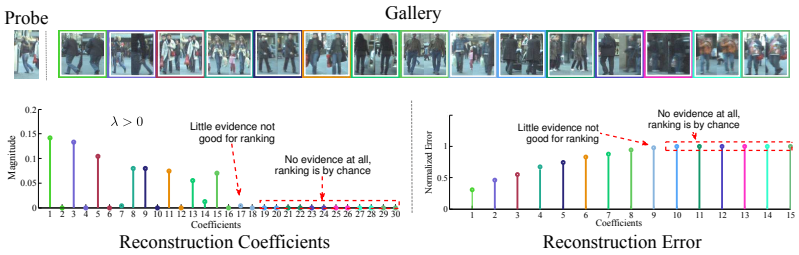


Figure 3.4: Ranking with limited information from a single basis expansion. A more detailed view of two problems that arise: ranking decisions must be made on the basis of little information (low coefficient energy) or *no* information (zero coefficient energy). Sparse reconstruction can only support ranking up to a certain rank for a non-zero value of λ. On the left are coefficients, on the right the corresponding normalized reconstruction error.

corresponding to the i-th subject is:

$$e_i = \frac{\|\mathbf{y} - \mathbf{T}_i \widehat{\boldsymbol{\alpha}}_{|i}\|_2}{\|\mathbf{y}\|_2}, \text{ for } i \in \{1, \ldots, C\}. \tag{3.7}$$

where $\widehat{\boldsymbol{\alpha}}_{|i}$ represents the sparse solution of Eq. (3.6) restricted to the coefficients corresponding to gallery examples of class i. That is, $\widehat{\boldsymbol{\alpha}}_{|i}$ is equal to $\widehat{\boldsymbol{\alpha}}$ at coefficients corresponding to gallery examples from individual i and zero elsewhere.

Our decision rule is:

$$\text{class}(\mathbf{y}) = \arg\min_i e_i. \tag{3.8}$$

This type of residual-based decision rule has been shown to be very effective for frontal face recognition [128].

3.4.2 Ranking with sparse basis expansions

The decision rule based on sparse discriminative basis expansion described in the previous section performs very well for classification problems. However, for re-identification it is also important to be able to rank the entire set of gallery individuals. We can extend the decision rule of Eq. (3.8) in a straightforward manner to rank candidate individuals using their corresponding residual e_i.

In Fig. 3.4 we show two views of ranking gallery individuals according to the normalized reconstruction error e_i with respect to the probe $\mathbf{y}$. On the left are illustrated the reconstruction coefficients $\widehat{\boldsymbol{\alpha}}$ of a probe image reconstruction in terms on a multi-shot gallery. On the right are illustrated the normalized reconstruction errors e_i corresponding to each gallery individual. The fundamental problem with using discriminative sparse basis expansions derived from solutions to problems like Eq. (3.6) is that, for many reasonable values of λ, we are deliberately forcing the majority of coefficients to zero. This severely limits the number of ranks the basis expansion can support. In Fig. 3.4 we see that after the first few individuals (ranks), the coefficient energy collapses and we have no more information upon which to base subsequent ranking decisions. The result is that beyond this point we cannot rank the remaining gallery individuals.

A more subtle problem is that in many cases we may be basing ranking decisions on inadequate evidence from the basis expansion. After the first eight individuals in Fig. 3.4, even before collapsing to zero, there is very little coefficient energy upon which to base individual ranking decisions. In the next section we introduce an iterative sparse basis expansion technique that addresses these problems of lack of sufficient ranking support in sparse reconstructions.

3.5 Ranking via iterative sparse re-weighting

In this section we develop an iterative technique to address the problems with applying sparse discriminative classifiers to ranking, arriving in the process at an algorithm that is able to robustly perform re-identification up to all ranks. Our approach is an iterative extension of the weighting described in [32] which we use to first re-weight basis vectors in the sparse solutions to Eq. (3.6) and arrive at a more robust solution that does not rely on basis vectors contributing little to the reconstruction. A similar weighting approach is then be used to proceed with ranking after damping the influence of basis vectors that have already contributed to ranking.

We first re-write Eq. (3.6), adding weights to the coefficients in the ℓ_1-regularized minimization problem:

$$\widehat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \|\mathbf{y} - \mathbf{T}\boldsymbol{\alpha}\|_2^2 + \lambda\|\mathrm{diag}(\mathbf{w})\boldsymbol{\alpha}\|_1, \tag{3.9}$$

where $\mathbf{w} = [w_1, \ldots, w_N]$ is a vector of positive weights, with $N = \sum_{i=1}^{C} n_i$

so that there is one $w_i > 0$ for each $\alpha_{i,j}$ in the basis expansion. Just like the unweighted counterpart in Eq. (3.6), this convex problem can also be efficiently solved as a linear program. The weights w_i are free parameters in the convex relaxation, whose values can be used to penalize or favor specific basis vectors in the regularized expansion.

3.5.1 Soft weighting for robust ranking

Our first refinement step is designed to address the problem of ranking individuals based on normalized reconstruction error e_i resulting from very low coefficient energy, as can be seen in detail in Fig. 3.4.

Assume we have computed sparse reconstruction coefficients $\widehat{\boldsymbol{\alpha}}$ for a given instance of a re-identification problem. The first refinement step we perform is a sort of *soft-weighting* that is used to remove those coefficients that weakly contribute to the reconstruction of the given test sample. At each iteration we weight each element in the basis according to:

$$w_{i,j} = \frac{1}{|\alpha_{i,j}| + \varepsilon} \quad \text{for } i \in \{1 \dots C\} \text{ and } j \in \{1 \dots n_i\}, \tag{3.10}$$

where ε is chosen to be slightly smaller than the minimum nonzero coefficient of $\widehat{\boldsymbol{\alpha}}$ to avoid division by zero and to not influence the solution with respect to the other coefficients. We then solve a weighted Lasso problem by weighting the regularization magnitudes using the $w_{i,j}$ defined above:

$$\widehat{\boldsymbol{\alpha}}' = \arg\min_{\boldsymbol{\alpha}} \|\mathbf{y} - \mathbf{T}\boldsymbol{\alpha}\|_2^2 + \lambda \sum_{i=1}^{C} \|\mathrm{diag}(\mathbf{w}_i)\boldsymbol{\alpha}_i\|_1, \tag{3.11}$$

where $\mathbf{w}_i = [w_{i,1}, w_{i,2}, \dots, w_{i,n_i}]$ are the weights from Eq. (3.10) corresponding to the basis vectors for individual i. The new solution $\widehat{\boldsymbol{\alpha}}'$ is the refined solution that is used to rank individuals with respect to probe $\mathbf{y}$.

Fig. 3.5(b) graphically illustrates this soft weighting procedure. The initial solution contains a few dominant coefficients that contribute most to the reconstruction of the probe image $\mathbf{y}$. It also contains a number of basis vectors that contribute very little to the overall reconstruction, as indicated by very small coefficients in the initial solution shown in Fig. 3.5(a). The refined solution using the weights from Eq. (3.10) is shown in Fig. 3.5(b). Note that this refinement eliminates small coefficients from the solution and redistributes their energy among more relevant basis vectors. It is not equivalent to a simple thresholding of coefficients.

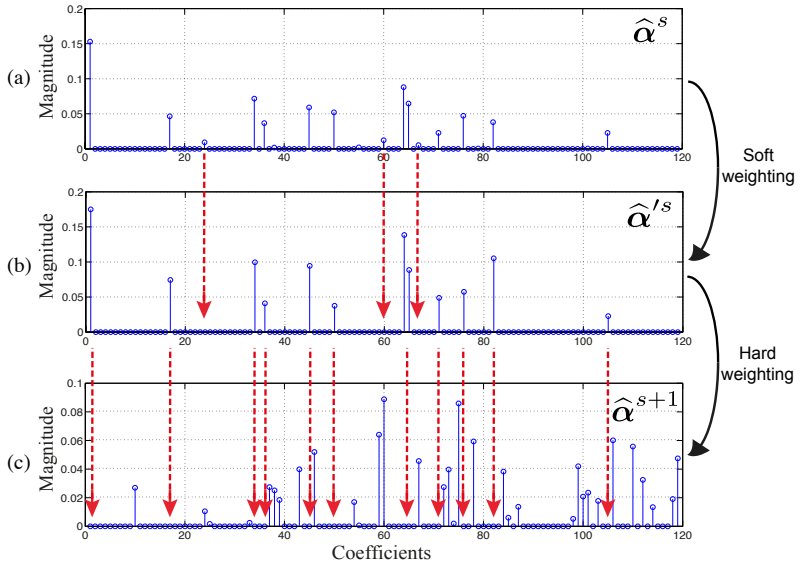


Figure 3.5: Soft and hard weighting of regularized reconstructions. (a) Reconstruction coefficients from the first solution $\widehat{\boldsymbol{\alpha}}^s$ at iteration s. (b) Refined reconstruction coefficients after soft weighting $\widehat{\boldsymbol{\alpha}}'^s$. (c) Coefficients $\widehat{\boldsymbol{\alpha}}^{s+1}$ at iteration $s+1$ after hard weighting.

3.5.2 Hard re-weighting for ranking completeness

A more serious problem related to using a sparse discriminative classifier for re-identification is the lack of sufficient non-zero support in sparse solutions. It is often the case that only a few gallery individuals can be ranked by analyzing the initial, refined sparse solution. On the VIPeR dataset, for example, it is common to report results up to rank 50 since it contains over 600 individuals. However, the sparse solution for $\lambda = 0.2$ only contains enough non-zero support to rank fourteen individuals (on average).

To address this problem we use a weighting scheme similar to the soft weighting described in the previous section. A set of hard weights are maintained that are used to exclude those elements that have already contributed to ranking an individual against the probe $\mathbf{y}$:

$$w_{i,j}^h \leftarrow \begin{cases} \infty & \text{if} \quad \alpha'_{i,j} > 0 \\ 1 & \text{otherwise} \end{cases} \tag{3.12}$$

Algorithm 1: Sparse iterative ranking

Input: $\mathbf{T} = [\mathbf{T}_1\ \mathbf{T}_2\ \cdots\ \mathbf{T}_C]$, the gallery templates;
$\mathbf{y}$, the probe template; and
λ, the regularization factor.

Output: R, the ranked gallery.

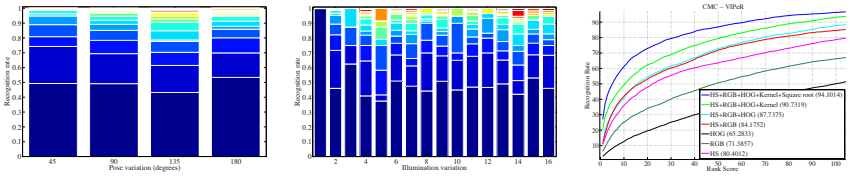
Initialize hard weights: $w^h_{i,j} \leftarrow 1$.
Initialize iteration count: $s \leftarrow 1$.
Initialize set of gallery individuals: $R \leftarrow \emptyset$.
while $|R| < C$ **do**
  $\widehat{\boldsymbol{\alpha}} \leftarrow \arg\min_{\boldsymbol{\alpha}} \|\mathbf{y} - \mathbf{T}\boldsymbol{\alpha}\|_2^2 + \lambda\|\mathrm{diag}(\mathbf{w}^h)\boldsymbol{\alpha}\|_1$
  $w_{i,j} \leftarrow \frac{1}{|\alpha_{i,j}|+\varepsilon}$ for $i \in \{1 \dots C\}$ and $j \in \{1 \dots n_i\}$
  $\widehat{\boldsymbol{\alpha}}' \leftarrow \arg\min_{\boldsymbol{\alpha}} \|\mathbf{y} - \mathbf{T}\boldsymbol{\alpha}\|_2^2 + \lambda\|\mathrm{diag}(\mathbf{w})\boldsymbol{\alpha}\|_1$
  for $\widehat{\boldsymbol{\alpha}}_i \neq \mathbf{0}$ **do**
    $R \leftarrow R \cup \{(i, s, e_i)\}$ (e_i from Eq. 3.7)
  end
  $s \leftarrow s + 1$
  $w^h_{i,j} \leftarrow \begin{cases} \infty & \text{if} \quad \alpha'_{i,j} > 0 \\ w^h_{i,j} & \text{otherwise} \end{cases}$
end
return R ordered by $(i, s, e) \leq (i', s', e')$ iff $s < s' \vee e \leq e'$

where $\alpha'_{i,j}$ are the coefficients from the soft-weighted solution $\widehat{\boldsymbol{\alpha}}'$. The hard weights vector $\mathbf{w}^h$ is used in the next step of an iterative process that repeats the soft weighting and ranking procedure. In Fig. 3.5(c) is shown the solution to the weighted Lasso problem using these weights. The difference in the distribution of coefficients between the hard weighted solution and the original solution in Fig. 3.5(a) is quite noticeable.

3.5.3 Ranking by iterative re-weighted ℓ_1 regularization

The process we use for person re-identification up to arbitrary ranks uses an iterative process of both soft- and hard-weighting of the ℓ_1 regularized reconstruction of probe images using gallery examples:

1. Reconstruct a probe image using Eq. (3.6).
2. Use *soft weighting* to eliminate coefficients that contribute little to the reconstruction of the probe and distribute their energy among more rel-



(a) Robustness to pose change (b) Robustness to illumination change (c) Contribution of each descriptor component

Figure 3.6: Descriptor robustness and the contribution of each descriptor component on the VIPeR dataset.

evant basis vectors.

3. Rank the gallery individuals who have non-zero energy in their corresponding coefficients (i.e. those individuals who have normalized reconstruction error $e_i < 1$).
4. Eliminate basis vectors contributing to ranking in the current round via hard weighting in subsequent iterations.
5. Repeat until all gallery individuals are ranked.

It is important to note that individuals are not ranked by their overall normalized reconstruction error alone. The normalized reconstruction error is used for ranking *within* iterations, but those individuals ranked in the first iteration will always be ranked higher than those in subsequent iterations. Algorithm 1 formalizes each of these steps performed and how they fit together to rank all individuals in the gallery.

Algorithm 1 is valid for both single- and multi-shot modalities. In the case of multi-shot modalities, a single gallery individual may appear more than once in the list of ranked results due to there being more than one exemplar in the gallery. In such cases, we use only the highest-ranked exemplar of each person and ignore any subsequent instances of that person in the ranked list of results.

3.6 Experimental Results

In this section we report results from an extensive set of experiments performed to evaluate our approach and to compare it with the state-of-the-art. All experiments were conducted on standard, publicly available datasets (VIPeR, i-LIDS, ETHZ and CAVIAR4REID), and we compare our results

with the following state-of-the-art approaches: SDALF [52], HPE [18], AHPE [19], SCR [15], ELF [62], CPS [34], MRCG [14], ContextB [140], PRDC [139], PRSVM [103], SBDR [129], EIML [71], COSMATI [13], RPLM [72] and eSDC [107].

We evaluate performance on all of the commonly used re-identification modalities found in the literature. The SvsS modality corresponds to when there are one or more (considered independently) exemplars for each person in the test set, and a single exemplar for each person in the gallery. The MvsS modality, on the other hand, is when there is a single exemplar of each person in the probe set, but one group of multiple exemplars for each person in the gallery. The MvsM modality is when there is a single group of multiple exemplars for each person in both the gallery and test sets.

The principal metric used for evaluating person re-identification is the Cumulative Match Characteristic (CMC) curve which summarizes overall performance by reporting recall over a range of cutoff points. A CMC curve represents the expectation of finding the correct match in the top r matches, where r is the rank considered in the final ranking result. For each experiment we split the dataset into probe and gallery sets for the various modalities. Each curve presented here was computed by averaging over 50 random, independent splits of dataset into gallery and probe sets.[3] We also report, and compare with the state-of-the-art when available, the normalized Area Under the Curve (nAUC). The nAUC is calculated as the total area under a CMC divided by $100 \times N$, where N is the total number of gallery individuals. It gives an overall score of how well a method performs over all ranks. For most applications, the most important cutoff rank is one (i.e. the first returned identity should be the correct one). We thus also report a comparison of our rank-1 performance compared to the state-of-the-art on all datasets.

In the next section we analyze a number of aspects of our approach, from the feature descriptor components to the contribution of our sparse approach with respect to baselines. Then in Sec. 3.6.2 through Sec. 3.6.5 we detail the performance of our algorithm with respect to the state-of-the-art on each dataset and for each modality. Finally in Sec. 3.6.6 we summarize our results and discuss computational requirements.

[3]Except for VIPeR where we use the ten splits from [52].

3.6.1 Descriptor robustness and sparse ranking

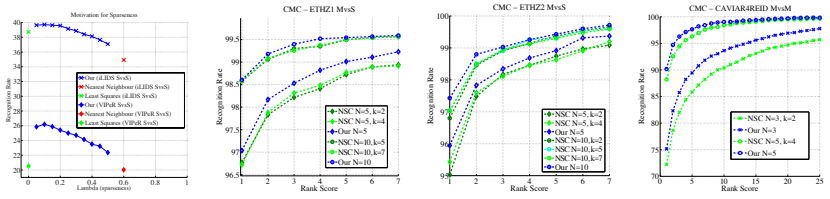
In this section we provide an analysis of each component of the descriptor proposed in Sec. 3.3. We also give some illustrative examples motivating the advantages iterated, sparse basis expansions have over other techniques.

Descriptor robustness: We performed an experiment on the VIPeR dataset to show the robustness of our descriptor to pose and illumination variations. We used this dataset since it provides pose information for each image. To estimate illumination variation for each pair of images we filtered the value channel in the HSV color space with a Gaussian kernel, weighted this filtered intensity channel with the Epanechnikov kernel, and then computed the average intensity. To evaluate how sensitive our descriptor and approach is to these variations we performed a leave-one-out experiment in which use each VIPeR image in sequence as a probe. We then recorded the rank at which the correct corresponding gallery image is returned as a function of pose and illumination variation.

In Fig. 3.6(a) and Fig. 3.6(b) each color represents 1/16-th of the possible ranks, ranging from dark blue for the first

The contribution of descriptor components: In Fig. 3.6(c) we report results from a series of experiments performed to evaluate the contribution of each component of our descriptor. These experiments were performed on the VIPeR dataset and averaged over ten trials. Results show that the addition of each component increases performance. We obtain the best results by adding the Epanechnikov kernel to weight the contribution of each pixel in the final HS and RGB histograms (in order to remove background information) and by taking the square root of each dimension of the descriptor.

The contribution of sparse ranking: In Fig. 3.7(a) we show the rank-1 re-identification accuracy for a variety of solutions of the regularized least squares problem of Eq. (3.6) over 10 trials on the VIPeR and i-LIDS datasets. Rank-1 accuracy is the percentage of probe images whose correct corresponding gallery image is ranked first by the re-identification algorithm. Shown are the least squares ($\lambda = 0$) solution, sparse solutions for a range of $\lambda > 0$, and the nearest neighbor solution when λ is sufficiently high to constrain the solution to a single non-zero coefficient. The sparse approach, for appropriate λ, outperforms the nearest neighbor and least squares solutions in terms of rank-1 accuracy. In fact, using our descriptor on the VIPeR dataset in combination with a nearest neighbor or least-squares approach, we arrive at



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

Figure 3.7: (a) Rank-1 re-identification accuracy on the VIPeR and i-LIDS datasets for the SvsS modality (10 trials). Accuracy is plotted for varying sparseness (λ), including the least squares solution when $\lambda = 0$ and the nearest neighbor solution for λ sufficiently large ($\lambda \approx 0.6$ for these datasets). Comparison of our sparse ranking approach with the Nearest Subspace Classifier on: (b) ETHZ1; (c) ETHZ2; (d) CAVIAR4REID. In the legend of each plot we report the number of instances per person (N) and the number of learnt subspaces (k).

a rank-1 recognition rate of only 20%, significantly lower than the state-of-the-art on VIPeR. We found $\lambda = 0.2$ to perform well and fixed this value for all subsequent experiments.

In Fig. 3.7(b,c,d) we provide a comparison of our approach with the Nearest Subspace Classifier (NSC) [89] on the MvsS and MvsM modalities on datasets with a large number of images per person (ETHZ1, ETHZ2 and CAVIAR4REID). We chose NSC as a baseline to motivate the use of a linear, sparse approach against a standard, non-sparse linear one. Briefly, the Nearest Subspace Classifier (NSC) learns a subspace for each gallery individual in feature space using PCA. Then, given one or more probes, it ranks the gallery considering the re-projection error over all the learnt subspaces. The drawback of the NSC is that it cannot robustly learn a subspace if the number of instances per person is low, while our method can robustly perform re-identification with as few as two or three gallery examples per person. We observe this in Fig 3.7(b,c) by how our approach outperforms NSC for small gallery sizes ($N = 5$), while the performance is more comparable when there are enough gallery images per person ($N = 10$) on which to learn more meaningful subspaces. The difference is especially noticeable in Fig. 3.7(d) on CAVIAR4REID, which has very poor image quality compared to ETHZ, when few images are used to learn subspaces.

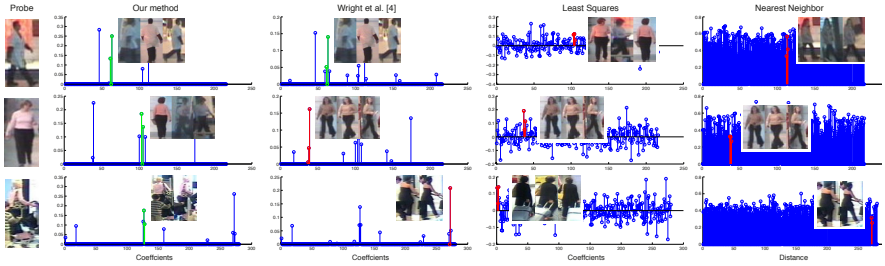


Figure 3.8: Examples of coefficient distribution for various techniques. Left to right: our approach, the method of [128], Least Squares and Nearest Neighbor. Examples are from the MvsS ($N = 3$) modality on i-LIDS and CAVIAR4REID datasets. Coefficients corresponding to true positives are in green while those in red correspond to false positives. Images on the plots show the rank-1 selected gallery set.

In Fig. 3.8 we show three example MvsS re-identifications on the i-LIDS and CAVAIR4REID datasets for $N = 3$. In the figure is given a comparison between our approach, Least Squares, Nearest Neighbor, and the technique of [128]. The first example demonstrates that methods like least squares and nearest neighbor are driven by the dominant color in the probe image. The other two examples demonstrate how our soft-weighting process re-distributes coefficient energy by adding more energy to the coefficients of other samples in the target basis. In this way our approach is able to leverage multiple aspects when more samples are available for each target.

3.6.2 Performance on the VIPeR dataset

The VIPeR dataset was the first standardized dataset proposed in the literature for the person re-identification task. It consists of 632 people imaged by two non-overlapping cameras. Image pairs exhibit viewpoint changes of up to 180 degrees and illumination changes that result in large intra-class variations. The dataset contains only two samples of each person (one from each view), and thus is only appropriate for the single-shot SvsS re-identification modality.

In order to fairly compare our results with other techniques, we use the publicly available splits into gallery and probe sets provided by the authors of SDALF [52]. Fig. 3.9(a) gives the CMC curves up to rank 50 comparing

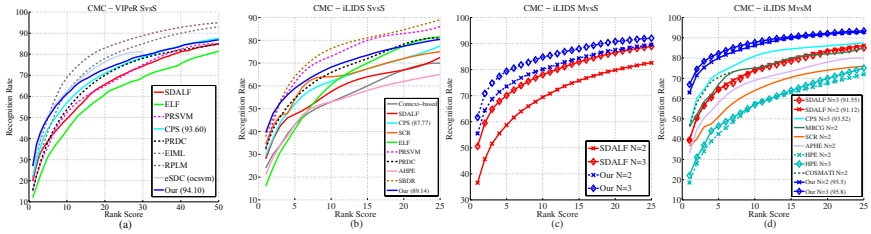


Figure 3.9: Comparative performance evaluation on VIPeR and i-LIDS (nAUC in parenthesis, when available). (a) The VIPeR SvsS modality. (b) The i-LIDS SvsS modality. (c) The i-LIDS MvsS ($N \in \{2, 3\}$) modality (d) The i-LIDS MvsM ($N \in \{2, 3\}$) modality. The dashed curves distinguish techniques that set aside a portion of the dataset for learning.

our approach with the many methods evaluated on VIPeR. We outperform all state-of-the-art techniques that are not based on metric learning up to all but the highest ranks. After about rank-5, techniques that learn on a part of the data like EIML [71], RPLM [72], and eSDC [107] begin to outperform our approach. Note that such techniques are not strictly comparable with ours since they set aside a portion (up to half) of the dataset on which to learn metrics. The gallery and probe sets are drawn from the remaining data and thus the standard splits cannot be used.

Table 3.1 gives the rank-1 performance of our approach with respect to the state-of-the-art on VIPeR. From this table we see that our method improves by about 6% over the state-of-the-art performance on the VIPeR dataset except for learning-based methods like RPLM [72] and eSDC [107], which perform similarly to ours at rank-1.

3.6.3 Performance on the i-LIDS dataset

The i-LIDS dataset was created by authors of [140] using images taken from the i-LIDS Multiple-Camera Tracking Scenario dataset of multiple camera views from a busy airport arrival hall. It contains 119 people with about four images of each person. As shown in Fig. 3.9(b), our approach outperforms the state-of-the-art at low ranks. After about rank-4, however, techniques based on metric learning begin to outperform us. Note that, due to having to use a portion of available data for learning, the SBDR [129] and PRSVM [103] methods only consider, respectively, 80 and 108 out of the 119 people in the

Methods ↓	**VIPeR**	**iLIDS**	
Modality →	*SvsS*	*SvsS*	*MvsM (2)*
HPE [18]	–	–	18.5
AHPE [19]	–	21	32
SCR [14, 15]	–	34.5	36
MRCG [14]	–	–	46
ContextB [140]	–	24	–
SDALF [52]	19.9	28	39
ELF [62]	12	16	–
CPS [34]	21.8	29.5	44
PRSVM* [103]	15	32	–
PRDC* [139]	15.7	32.6	–
SBDR* [129]	–	37.75	–
EIML* [71]	22.0	–	–
COSMATI* [13]	–	–	44
eSCD (ocsvm)* [107]	26.7	–	–
RPLM* [72]	**27.0**	–	–
Our	**27.0**	**39.5**	**62.9**

Table 3.1: Comparative performance analysis at rank-1 with respect to the state-of-the-art on VIPeR and iLIDS. Techniques indicated with a "*" set aside a portion of the dataset for learning and splits are generated from the remaining data.

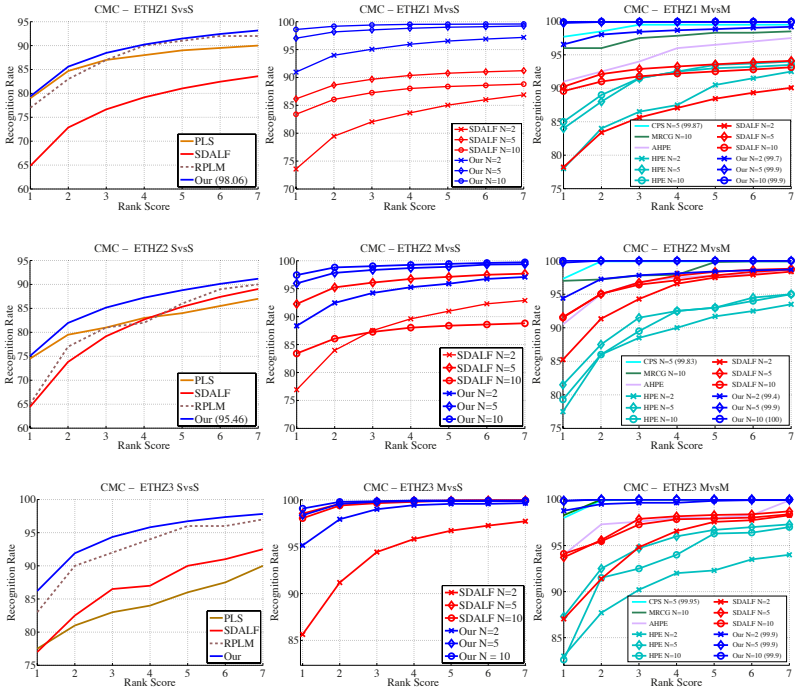


Figure 3.10: Performance of our approach on the ETHZ dataset compared to state-of-the-art methods. The first row shows results on ETHZ1, the second row ETHZ2, and the third row ETHZ3. Each column represents a re-identification modality: the first column SvsS, the second column MvsS for $N \in \{2, 5, 10\}$, and the third column MvsM for $N \in \{2, 5, 10\}$.

dataset.

In the MvsS and MvsM modalities, where our technique is able to exploit multiple images of each gallery individual, our improvement over the state-of-the-art is even more dramatic. In particular, as seen in Fig. 3.9(c) for MvsS ($N = 2$) we exceed the state-of-the-art at rank-1 by nearly 20%. We similarly improve for MvsS ($N = 3$) where we outperform SDALF by nearly 15% at rank-1. For the MvsM case we report results for $N \in \{2, 3\}$ in Fig. 3.9(d) along with results of other methods tested on this dataset. We outperform the state-of-the-art at all ranks for the MvsM scenario.

Table 3.1 also summarizes the rank-1 performance of our technique with

Methods ↓	ETHZ1			ETHZ2			ETHZ3			CAVIAR4REID		
Modality →	*SvsS*	*MvsM (5)*	*MvsM (10)*	*SvsS*	*MvsM (5)*	*MvsM (10)*	*SvsS*	*MvsM (5)*	*MvsM (10)*	*SvsS*	*MvsM (3)*	*MvsM (5)*
HPE [18]	–	84	85	–	81.5	79.3	–	87.3	82.6	–	–	–
AHPE [19]	–	91	–	–	90.6	–	–	94	–	7.5	7.5	7.5
MRCG [14]	–	–	96	–	–	97	–	–	98.3	–	–	–
PLS [109]	79	–	–	74.5	–	–	77.5	–	–	–	–	–
SDALF [52]	64.8	90.2	89.6	64.4	91.6	91.5	77	93.7	94.1	7	8.5	8.3
CPS [34]	–	97.7	–	–	97.3	–	–	98	–	8.5	13	17.5
EIML* [71]	78	–	–	74	–	–	**91**	–	–	–	–	–
RPLM* [72]	77	–	–	65	–	–	83	–	–	–	–	–
eSDC* [107]	**80**	–	–	**80**	–	–	89	–	–	–	–	–
Our	79.5	**99.8**	**99.9**	76.1	**99.7**	**100**	86.2	**99.9**	**99.9**	**29**	**75.1**	**90.1**

Table 3.2: Comparative performance analysis at rank-1 with respect to the state-of-the-art on the ETHZ and CAVIAR4REID datasets. The EIML, RPLM, and eSDC techniques (indicated by "*") set aside a portion of the data for metric learning.

respect to the state-of-the-art for the SvsS and MvsM modalities on the i-LIDS dataset. From this table we see that we slightly outperform other approaches on the SvsS modality, while we significantly outperform competing methods by about 20% for MvsM.

3.6.4 Performance on the ETH Zurich datasets

The ETH Zurich dataset consists of three sequences used for tracking, from which Schwartz and Davis [109] extracted a set of samples of each person in the videos. They extracted 83 people from the first sequence (ETHZ1), 35 from the second (ETHZ2), and 28 from the third (ETHZ3). For each sequence they identified about 10 images per person. We performed SvsS, MvsS and MvsM experiments, varying the number of elements in both the probe and the gallery.

As can be seen in Fig. 3.10, the recognition rate of our method and state-of-the-art methods is very high (about 75-80% at first rank) for each modality. This is due to the fact that images in both the probe and gallery sets are taken from tracking sequences and thus from the same camera. In this figure, each row represents one sequence of the dataset, while each column represents a modality. The first column reports the SvsS case, the second reports results for the MvsS modality for $N \in \{2, 5, 10\}$, and the third reports the MvsM case for $N \in \{2, 5, 10\}$. Our approach outperforms current methods in all three modalities, with the improvement particularly acute in multi-shot scenarios MvsS and MvsM. In Table 3.2 we also report rank-1 results for each sequence of the ETHZ dataset for the SvsS and MvsM

($N \in \{5, 10\}$) modalities. Results on these datasets are quite saturated.

3.6.5 Performance on the CAVIAR4REID dataset

The CAVIAR4REID dataset[4] contains 72 unique individuals captured in a shopping center scenario: 50 are extracted from two camera views (and so are imaged from two distinct, but overlapping cameras), while the remaining 22 are extracted from a single view. This dataset was designed to maximize variability with respect to resolution changes, illumination conditions, occlusions, and pose changes.

In Fig. 3.11(a) we report the CMC curves for our approach and the state-of-the-art on the SvsS modality. Our approach outperforms current methods up to about rank-20. The improvement over the state-of-the-art at first rank is particularly noticeable: there is a difference of 22.5% at rank-1 between our performance and competing methods. In the legend we also report the nAUC for each method, which gives an idea of the trend of the curve across all ranks. Fig. 3.11(b) reports CMC curves for the MvsM modality on the CAVIAR4REID dataset for $N \in \{3, 5\}$. In the MvsM modality, as for i-LIDS, we significantly outperform the state-of-the-art at all ranks.

We report the rank-1 recognition rate of our approach on CAVIAR4REID compared to the state-of-the-art in Table 3.2. Our approach significantly outperforms competing methods at rank-1 in all modalities on this dataset. On the MvsM modality we improve on the state-of-the-art by nearly 62% for MvsM ($N = 3$) and by 73% for MvsM ($N = 5$).

3.6.6 Discussion

In this section we summarize the contributions of our approach in terms of performance with respect to the state-of-the-art and in terms of computational efficiency.

General performance considerations: The trend that emerges from the experimental evaluation in the previous sections is that, generally, our approach exceeds the state-of-the-art in terms of rank-1 accuracy. This can be seen in the SvsS modality on all datasets, but the increase in performance is particularly noticeable in the multi-shot MvsS and MvsM modalities on i-LIDS, ETHZ and CAVIAR4REID. In these multi-shot cases our approach exceeds the state-of-the-art at all ranks on these three datasets.

[4] `http://www.lorisbazzani.info/code-datasets/caviar4reid/`

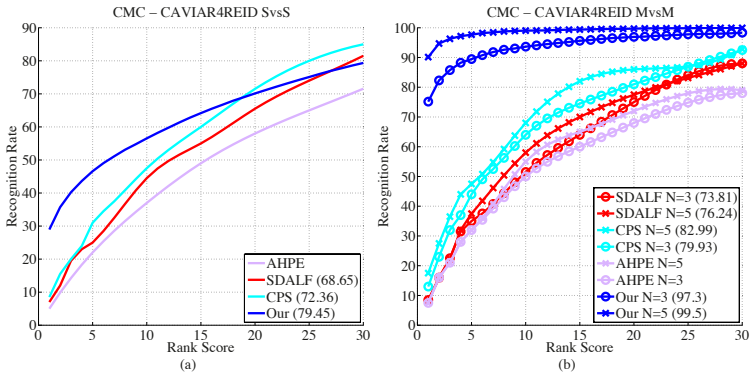
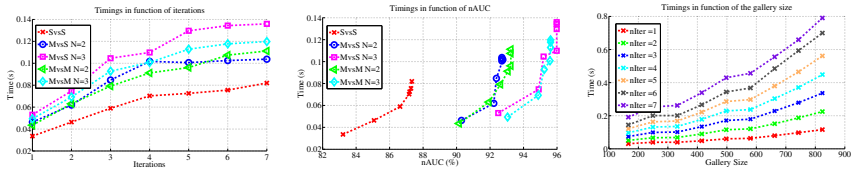


Figure 3.11: Performance on the CAVIAR4REID dataset with respect to the state-of-the-art. (a) The SvsS modality. (b) MvsM for $N \in \{3, 5\}$. In the legends we report the normalized area under the CMC curve (nAUC), when available.

Our method fully exploits the robustness of our descriptor with sparse reconstruction in order to achieve highest performance at the rank-1 for all scenarios, including SvsS. Our ranking based on sparse, ℓ_1-regularized basis expansions allows our approach to exploit multiple aspects of person appearance in the multi-image gallery sets of the MvsS and MvsM modalities. This can be appreciated from the trend of the curve on the i-LIDS dataset in Fig. 3.9(d) where we quickly reach a 90% recognition rate at around rank-15 and on CAVIAR4REID in Fig. 3.11(b) where we reach 100% accuracy around rank-20. On the ETHZ datasets we reach between 99.7% to 100% in the first seven ranks as shown in Fig. 3.10.

Comparison with metric learning: Some metric learning approaches outperform our method at higher ranks for the SvsS modality on the VIPeR and i-LIDS datasets. By setting aside a portion of the labeled data they are able to learn a metric that better captures the intrinsic properties of the scene, of the cameras used, and of the camera positioning and imaging conditions. This increase in high rank performance comes at a cost, however. On VIPeR, for example, as much as half of all the available labeled data (that is 316 image pairs) is used for metric learning and this limits the availability of data for actual testing. Not only does this render experimental results not strictly comparable, it is also a severe limitation in real application scenarios where no labeled data may be available *a priori*. An important advantage



(a) Time as a function of iterations.

(b) Time as a function of desired nAUC.

(c) Time as a function of gallery size.

Figure 3.12: Time required for re-identification. (a) Time as a function of the number of iterations of sparse re-weighted ranking on the i-LIDS dataset. (b) Time for a desired nAUC on the i-LIDS dataset. (c) Time as a function of the total number of gallery images on the ETHZ1 dataset for the MvsS $N = 10$ modality. Timings are averaged over all probes in fifty random splits.

of our method with respect to learning-based ones is that learned distance metrics cannot be easily updated if the camera settings or positions change, while for our approach it is straightforward to integrate new instances per person and to discard old gallery images.

We also note that our approach is complementary to metric learning. Our core algorithm for iterative sparse ranking is feature agnostic and could be applied to any underlying feature space, including those induced by learned metrics like the one proposed in [72]. We expect additional improvements from such a combination.

Computational efficiency: Our approach is implemented in MATLAB and makes extensive use of the optimized SPAMS library for sparse modeling [92]. All tests were performed on an Intel Xeon@2.67GHz (8-core) with 12 GB RAM.[5] Descriptor extraction in MATLAB requires about 0.016s per person image and is included in all timing numbers reported here.

In Fig. 3.12 we report three views of the computational requirements of our approach. In Fig. 3.12(a) we vary the number of iterations of sparse re-weighted ranking we perform in order to quantify how computational requirements change with increasing iterations (and increasing accuracy at higher ranks). Fig. 3.12(b), on the other hand, quantifies the relationship between the time required for performing a single re-identification and the area under

[5]The source code for our approach is available at `http://www.micc.unifi.it/lisanti/source-code/re-id/`

the curve. From these curves we can observe that, if we are interested only in first rank, with our approach we can perform re-identification of a single probe in about 0.036s. In real application scenarios our approach can thus perform rank-1 SvsS person re-identification at about 30 re-identifications per second.

If we are interested higher ranks, for example in an interactive application in which a human operator will sift through re-identification results, our approach might require more than one iteration. From Fig. 3.12(b) we see that after 7 iterations we arrive at a nAUC of about 88%, requiring 0.08s to compute this result (which works out to about 12 re-identifications per second). In the MvsM modality our approach requires about 0.14s (7 re-identifications per second), but yields a nAUC of more than 94%. This first two tests are carried out on the i-LIDS dataset.

In Fig. 3.12(c) we show how our approach scales as a function of the gallery size. In particular, we see that the time for a single re-identification increases approximately linearly when increasing the number of images in the basis up to 600 images; then the trend becomes superlinear from 600 to 900. It is interesting that this non-linearity is more pronounced with increasing number of iterations. This test was carried out on the ETHZ1 dataset which contains the biggest number of images. All the timing measurements are obtained by averaging over all probe samples from 50 random splits of gallery/probe image sets.

3.7 Conclusions

In this chapter we described an approach to person re-identification that is based on sparse, ℓ_1 regularized basis expansions of probe images in terms of gallery images used as basis vectors. We showed how to extend, through iteration and re-weighting, the concept of a Sparse Discriminative Classifier to problems requiring ranked output. Our algorithm is efficient and obtains state-of-the-art performance on both multi- and single-shot person re-identification modalities. Our results demonstrate how sparse reconstruction generally leads to higher performance at first rank, while also yielding higher nAUC using the proposed iterative ranking. Our method is also competitive with respect to metric learning based methods which set aside a portion of available data for training.

Our approach makes use of a simple, yet discriminative descriptor of per-

son appearance. It requires no foreground/background separation or body part segmentation. It is simple and extremely efficient to calculate, and the performance of our approach demonstrates that simple descriptors can be successfully applied in re-identification scenarios.

Iterative sparse ranking is a general approach and can be applied to retrieval problems beyond person re-identification. Additionally, we note that, though all of the experiments in this article were performed using the proposed descriptor, our sparse basis expansion approach is complementary to the feature representation used. It is feature agnostic and it can be similarly applied to any feature that is encoded as a fixed-length vector.

Part II

Face Analysis

Chapter 4

Logging of Face Imagery

Faces, particularly high-quality imagery of faces, can be a powerful semantic cue for forensic analysis of video footage. In this chapter we present a real-time solution for posterity logging of face images in video streams. Our system detects and tracks multiple targets in real time, grabbing face images and evaluating their quality to store only the best for each detected target. We propose two quality measures for face imagery, one based on symmetry and the other on face pose. Extensive qualitative and quantitative evaluation of the performance of our system is provided on many hours of realistic surveillance sequences.[1]

4.1 Introduction and Related Work

Modern surveillance systems are designed to record what happens, twenty four hours per day, every day. These video streams must be laboriously and manually searched by forensic investigators for evidence in the case of an incident. The result of this is enormous quantities of archived surveillance video that is largely inaccessible due to the lack of high-level, semantically meaningful annotations. What is needed are tools for filtering through this massive amount of video for data, events and features that *might* be useful in the future for the purposes of forensic reconstruction of an event. To this end, a log of interesting events, or a *posterity log* of interesting details, would be

[1]This chapter has been published as "Posterity Logging of Face Imagery for Video Surveillance" in the journal *IEEE Multimedia*, October 2012.

a tool of great utility to forensic investigators engaged in the post mortem analysis of video events. Face logging, in particular, can unambiguously associate an "identity" in the form of a high-resolution face image with each individual entering an area of surveillance.

The main operating principle of surveillance systems does not change: cameras continue to record and archive multiple streams of video, but these streams are now augmented with a log annotating the faces of individuals entering the scene. These logs should satisfy three primary requirements to be useful to forensic investigators:

- **logs should be brief** and summarize the event in a convenient format that does not overload downstream investigators with redundant images of the same target;
- **logs should be high-quality** and provide some guarantee about the quality of face imagery they extract; and
- **logs should be accurate** and provide the minimum amount of false positives (i.e. logs not corresponding to faces) as well as false negatives (i.e. faces not logged).

A few works address the problem of face tracking in video with a focus on image quality. The work of Fronthaler et al. [58] is focused on image quality estimation using the orientation tensor with a set of symmetry descriptors to estimate the quality of acquired fingerprints or faces. Subasic et al. [116] propose a quality evaluation technique that checks several low and high level image parameters, according to the requirements proposed by the International Civil Aviation Organization.

To the best of our knowledge, Fourney and Laganière [57], and Nasrollahi and Moeslund [96] are the only works explicitly addressing the face logging problem by building sequences of increasing quality face images. In particular, the technique described in [96] is similar to the system proposed by Fourney and Laganière, but instead of a linear combination uses a fuzzy combination of primitive quality measures. Note that the main focus of their approach is to establish a measure of facial image quality given a set of face images, and not to describe an entire face logging system performing detection, multi-target tracking and evaluation of facial image quality. Neither approach has been evaluated on significant quantities of realistic surveillance video footage.

The novel aspects of our contribution are:

- the development of a multi-pose face detector, based on the AdaBoost

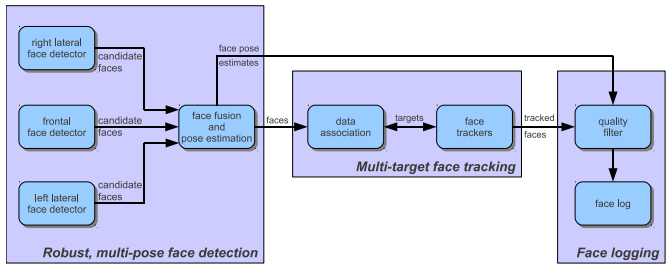


Figure 4.1: A schematic diagram of our proposed system. A robust, multi-pose face detector feeds multi-target tracking and data association modules, and a quality filter ensures compact, high-quality logs are saved in the final stage.

face detector of Viola and Jones [87, 123], that incorporates the outputs of lateral and frontal face detectors to improve robustness to face pose variations and provides a rough estimate of face pose usable as a facial quality measure;

- the definition of a multi-target tracking and data association module handling multiple intrusions over long periods of time;
- the proposal of two distinct facial quality measures, in addition to resolution, which are compared against the brute force approach that logs each detection associated with a track.

The integration of these novel components, along with state-of-the-art components for object detection and tracking, outperforms other face logging methods described in the literature. An extensive evaluation of our face logger on more than ten hours of diverse and realistic surveillance videos, with both quantitative and qualitative analysis, is provided. These videos have been released as a new publicly available dataset to support and encourage comparative evaluation of face logging systems. The dataset is available at `http://www.micc.unifi.it/datasets/logging-of-face-imagery/` and contains fourteen sequences of more than ten hours of videos and 74,871 annotated faces. For more details about the dataset please see Appendix A.

4.2 The face logging system

Our face logging system is designed using a pipelined architecture (see figure 4.1) where each frame is passed through a pipeline of several stages, each devoted to a specific function. The pipeline is composed of the following components:

- the **multi-pose face detectors** are responsible for detecting all targets in newly grabbed frames (described in section 4.2.1);
- **face fusion** is then used to fuse the results of the multi-pose detectors into confidence-scored detections with a rough estimate of face pose (also described in 4.2.1);
- **face trackers** based on a particle filter are responsible for the tracking of each target in a sequence (described in section 4.2.2);
- **data association** is responsible for matching the set of detected faces with the set of targets already being tracked (described in section 4.2.3); and
- finally the **quality filter** evaluates the quality of extracted faces in order to build a reliable log of each target (described in section 4.2.4).

In the following sections we describe each component of our system in detail.

4.2.1 Multi-pose face detection

The face detection stage plays a critical role in the system since its results are used for tracker initialization, the measurement process which keeps trackers fixed on their targets, and quality evaluation for the construction of logs. The detection approach we propose exploits the well known AdaBoost technique [123] to implement a robust detector capable of detecting not only frontal faces, but left and right lateral profile ones as well. Differently from [43], we use the outputs of three independent detectors. Most other multi-pose detectors use a single cascade trained for multi-pose detection. For example, [131] concentrates on robustness to in-plane rotations and expansion of the Haar basis to improve multi-pose detection using a single cascade, while [106] uses skin color and landmark to pre-filter and group face candidates with similar poses. Our approach allows us to fully exploit modern, multi-core architectures and to extract a facial quality measure based on how "close-to-frontal" a candidate face is.

Given that detector performance gracefully degrades in the neighborhood of the face location, each component classifier outputs more than one hypoth-

esis for an object. Considering this, for each putative detection a weighted confidence is obtained as a linear combination of the number of neighbor hypotheses suppressed through non-maximum suppression.

The N detection responses from all three detectors are collected together into a set $\mathbf{D} = \{\mathbf{d}_i\}_{i=1}^{N}$. Each detection is represented by its bounding box geometry and a vector indicating which type and number of cascade responses it represents:

$$\mathbf{d}_i = \{\mathbf{b}, \mathbf{n}\}, \quad i = 1 \dots N, \tag{4.1}$$

where $\mathbf{b} = [x, y, w, h]^{\mathrm{T}}$ represents the geometry of the bounding box of the detection, and $\mathbf{n} = [n_f, n_l, n_r]^{\mathrm{T}}$ where exactly one of n_f, n_l, n_r is non-zero and indicates whether detection $\mathbf{d}_i$ represents a frontal, left lateral or right-lateral detection and the number of cascade responses merged through non maximum suppression.

Our approach then merges the candidate detections in $\mathbf{D}$ that are spatially overlapping, and thus come from different detectors. In the end, each $\mathbf{n}$ will be a histogram counting the number of frontal, left lateral and right lateral detections that overlapped with $\mathbf{b}$ before being suppressed by non-maximum suppression in the AdaBoost detection algorithm. We iteratively select the two nearest detections in $\mathbf{D}$. If these two detections, denoted by $\mathbf{d} = \{\mathbf{b}, \mathbf{n}\}$ and $\mathbf{d}' = \{\mathbf{b}', \mathbf{n}'\}$, overlap by at least 70%, they are removed from $\mathbf{D}$ and replaced with a new detection having a rectangle corresponding to a weighted average of $\mathbf{b}$ and $\mathbf{b}'$. Letting $n = \mathbf{n}^{\mathrm{T}}\mathbf{1}$ and $n' = \mathbf{n}'^{\mathrm{T}}\mathbf{1}$, the new detection is:

$$\mathbf{d}_* = \left\{ \frac{n\mathbf{b} + n'\mathbf{b}'}{n + n'}, \mathbf{n} + \mathbf{n}' \right\}. \tag{4.2}$$

In this way, each original detection rectangle is weighted by its confidence as represented by the number of detections contributing to it. This procedure is iterated until the two nearest detections in $\mathbf{D}$ overlap by less than 70%. The resulting detections contain, in the vector of detector response counts, a measure of how confident each of the component detectors was at that point in the image.

To obtain a rough estimate of a 1D face pose $\bar{\phi}$ from the detector responses we exploit the number of candidate neighbors $\mathbf{n}_i$ using a set of weights $\mathbf{p} = [\alpha, \beta, \gamma]^{\mathrm{T}}$ and consider angles corresponding to the three poses of the detectors (determined by the training data used for training each detector) $\boldsymbol{\phi} = \{\phi_f, \phi_l, \phi_r\}$. Given a merged detection $\mathbf{d}_* = \{\mathbf{b}_*, \mathbf{n}_*\}$, where

$\mathbf{n}_* = [n_f, n_l, n_r]^{\mathrm{T}}$, the pose is estimated as:

$$\bar{\phi} = \frac{\boldsymbol{\phi}^{\mathrm{T}}[\alpha n_f, \beta n_l, \gamma n_r]}{\mathbf{p}^{\mathrm{T}}\mathbf{n}_*}. \tag{4.3}$$

For the frontal, left lateral and right lateral detectors we use the corresponding angles $\phi_f = 2\pi$, $\phi_l = -\frac{\pi}{2}$, $\phi_r = \frac{\pi}{2}$. The weights used are $\alpha = 1.3$, $\beta = \gamma = 0.5$ and were determined by empirical evaluation. Once the detections are merged and the pose estimated, a detection is signaled if and only if its confidence $c_* = \mathbf{p}^{\mathrm{T}}\mathbf{n}_*$ is greater than 10.

4.2.2 Face trackers

A multiple face tracker must be robust to partial and complete occlusions, to appearance changes due to changing expression, to illumination changes and noisy, chaotic scenes in general.

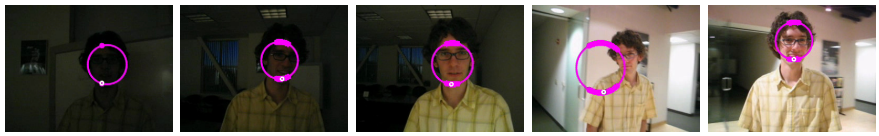
Seq	# of frames	Adaptive Particle Filter (ours) [42]	CAM-shift tracker [5]	Zhou's tracker [141]	Stalder's tracker [115]
1	96	79 (82.2%)	**90** (**93.7**%)	3 (3.1%)	10 (10.4%)
2	49	33 (67,3%)	**43** (**87.7**%)	8 (16.3%)	22 (44.8%)
3	191	185 (96.8%)	**191**(**100.0**%)	47 (24.6%)	47 (24.6%)
4	116	**115** (**99.1**%)	114 (98.2%)	91 (78.4%)	104 (89.6%)
5	115	**86** (**74.7**%)	3 (2.6%)	38 (33.0%)	34 (29.5%)
6	97	**89** (**91.7**%)	1 (1.0%)	39 (40.2%)	53 (54.6%)
7	1283	1204 (93.8%)	**1272** (**99.1**%)	6 (0.4%)	56 (4.3%)
8	937	**925** (**98.7**%)	2 (0.2%)	87 (9.2%)	262 (27.9%)
9	933	**794** (**85.1**%)	578 (61.9%)	12 (1.2%)	13 (1.3%)
10	1188	**761** (**64.0**%)	321 (27.0%)	4 (0.3%)	52 (4.3%)

Table 4.1: Comparative performance analysis of the tracker we use against state-of-the art tracking methods (table from [42]). These sequences test the trackers over a range of common, difficult situations: full and partial occlusion, out of plane rotations, long sequences and aspect ratio changes. Performance is measured by tracking success rate and the best results on each sequence are highlighted in bold.

Due to the difficult nature of face tracking in real scenes, careful uncertainty estimation is needed. The tracker we use as a basis for multiple face tracking is described in detail in [42]. An adaptive particle filter–based algorithm tracks position and velocity of each target using a normalized color histogram to describe target appearance. Measurements are provided to each particle filter from detections resulting from the robust, multi-pose face

(a) Tracker from [105] on a sequence from Dataset #1.

(b) Our face tracker on a sequence from Dataset #1.



(c) Our face tracker on David Indoor sequence from [105].

Figure 4.2: A comparison of our face tracker with the one described in [105]. We superimpose on the tracked face a circle with an indicator that shows the responses of the three learned detectors that coarsely indicate the face pose of the subject.

detector described in the previous section. When detections are not available the normalized color histograms are used due to their scale invariance and robustness to illumination changes. Table 4.1 reproduces some of the results of the comparative evaluation performed in [42]. Though the CAM-shift tracker [5] sometimes outperforms the adaptive particle filter on short sequences, the noise-adaptive capability of our tracker yields consistently better tracking results on longer sequences. Robust tracking over long sequences is essential for face logging in order to extract the best quality face imagery possible, to associate it with the correct log and to avoid creating new logs from broken tracks.

To determine its suitability for tracking multiple faces, we conducted a number of experiments with another state-of-the-art tracking-by-detection tracker [105] (shown in figure 4.2). In this figure we illustrate how each tracker performs in critical situations occurring frequently in face logging

scenarios. For the comparison we use one of our videos with a single intrusion from the first dataset and the David Indoor sequence from [105], in which a person moves from a dark to a bright area, undergoing large illumination and pose changes. We can see in figure 4.2(a) that the tracker of [105] on our sequences drifts after a few frames and remains fixated on the door in the background. Our tracker performs very well both on our sequence, as shown in figure 4.2(b), and on the David Indoor sequence from [105], as shown in figure 4.2(c).

4.2.3 Data association

The data association module is a critical component of the face logging application, as it associates trackers and their target's measurements in video frames. Firstly, it makes it possible to assign a detected face to an already running tracker. Secondly, it allows the face logger to deal with the target birth process. At every frame, the position of each target is predicted one time step ahead, and the result is compared with the set of all detections obtained on that frame. A matrix of the score of each tracker against each candidate face is then computed using the VOCscore criterion:

$$\text{VOCscore}(T_i, F_j) = \frac{|T_i \cap F_j|}{|T_i \cup F_j|}, \tag{4.4}$$

where T_i and F_j respectively represent a bounding box of a tracked target and a detected face and $|\cdot|$ is the area of the resulting rectangle. Once this matrix is available, data association proceeds iteratively by selecting the maximum non-zero value and removing the corresponding tracker–face pair from the pool. This process continues until all trackers have been assigned a face or all faces have been assigned to a tracker, whichever comes first. If a face was not assigned to any tracker, a new tracker is initialized with that face as a target.

The data association module also handles target death by implementing a number of criteria to decide whether or not a tracker should be stopped. The first is based on a measure of tracking quality estimated from histogram similarity as described in [42]. A tracker that is not capable of accurately locating its target any more is stopped. If the lost target is detected again, it will be assigned a new identity and a new tracker will be started (and consequently a new log created). The second criterion concerns the number of iterations elapsed since the last time the tracker received a measurement

from the our multi-pose face detector. By counting the tracker iterations since the last received face observation, the system can decide to stop a tracker which has not received information from the face detector for too long. The final death criterion handles the situation in which two trackers collide. In this case, the short distance between targets may lead to undesirable consequences such as an identity mismatch or identity switch. In order to reduce the probability of this, when a collision occurs the system stops the tracker that has not received measures from the face detector for the longest time.

4.2.4 Quality filters

The concept of image quality is strongly dependent on the purpose an image is intended for. In [57] and [96], face quality estimation is performed using measures such as sharpness, illumination, image resolution and pose. Sharpness and illumination can be subjective indicators of facial image quality and additionally require careful tuning to accommodate varying illumination and background clutter conditions. In our experience, the AdaBoost face detector, and subsequently our multi-pose face detector based on it, deliver reasonably sharp imagery across a range of illumination conditions when supplied with high-quality face images when training the cascades. Thus we propose propose here two quality measures based on image symmetry and face pose to aid in the filtering of face images.

For each detection associated with a tracked target that passes the stage described in section 4.2.1, three quality scores are computed: one for resolution, one for symmetry and one for face pose. Symmetry and face pose are redundant measures of face image quality in that they are highly correlated and using both has little sense. Symmetry has the advantage of being efficient to compute, while our face pose measure results practically as a side effect of our multi-pose face detector described in section 4.2.1.

The resolution score $\mathbf{R} \in [0, 1]$ is computed as the ratio between the face image size and the overall frame size: $\mathbf{R} = \frac{|F|}{|I|}$, where F and I denote a face sub-image and the whole frame, respectively.

The symmetry score $\mathbf{S}_{sym}$ is computed as the residual power of the signal obtained by differencing the hue channel of the original image with that of the mirrored image. The result is normalized in order to obtain a score in

$[0, 1]$:

$$\mathbf{S}_{sym} = 1 - \frac{\sum_{x=0}^{w-1} \sum_{y=0}^{h-1} |F_H(x, y) - F_H(w - x - 1, y)|}{|F|}, \tag{4.5}$$

where F_H denotes the hue channel of a candidate face sub-image. The symmetry score $\mathbf{S}_{sym}$ equals 1 when F is perfectly symmetric, and equals 0 when antisymmetric. We found this quality measure to be simple and effective, and further to be robust across a variety of illumination conditions. Only in cases where images are highly over- or under-saturated does this measure degrade significantly.

A "closeness to frontal" quality measure, $\mathbf{S}_{pose} \in [0, 1]$, can be computed from the pose $\bar{\phi}_i$ of a detection $\mathbf{d}_i$ received from our detector described in section 4.2.1. This score can then be used to derive a weighting factor for the candidate face:

$$\mathbf{S}_{pose} = \exp\Big(-\frac{|\bar{\phi}_i - \phi_f|}{\sigma_{pose}}\Big), \tag{4.6}$$

where ϕ_f is the frontal angle 2π and σ_{pose} is a parameter that scales the tolerance for determining how close a face should be to frontal (set to 1 in all of our experiments).

In the next section we evaluate the facelogger using resolution in combination with these two facial image quality measures: resolution with symmetry as in eq. 4.7, and resolution with face pose as in eq. 4.8:

$$\mathbf{Q}_{sym} = W_R \cdot \mathbf{R} + W_S \cdot \mathbf{S}_{sym}, \quad W_R + W_S = 1 \tag{4.7}$$

$$\mathbf{Q}_{pose} = W_R \cdot \mathbf{R} + W_S \cdot \mathbf{S}_{pose}, \quad W_R + W_S = 1. \tag{4.8}$$

When varying W_R (and consequently $W_S = 1 - W_R$) in preliminary experiments we noticed that for values of W_R in the range $[0.1, 0.9]$ the system consistently returned nearly the same number of false positives. Near the extrema of the range (e.g. $W_R = 0$ and $W_R = 1$) there was, however, strong variation in the number of false positives. When using only W_R the system is more likely to accept a candidate since it considers only resolution, and indeed the number of false positive is higher; when using only W_S the system is less likely to accept a candidate since it is more sensitive to pose and thus generates fewer false positives. Considering this we chose a value in the middle of this range ($W_R = W_S = 0.5$) as a trade-off between false positives and resolution. If the quality of a candidate face image exceeds the

highest quality seen by the tracker, it is saved in the log associated with the tracker, otherwise it is discarded.

4.3 Experiments

We tested our face logging system on three datasets (16 sequences for a total of about 12 hours). These datasets consist of a varying number of sequences taken at a frame rate of about 20 fps at different resolutions (see figure A.2 in the Appendix).

The sequences in Dataset #1 were recorded to cover a wide variety of circumstances and to give a qualitative impression of performance. For example, some of them are taken from a typical door or corridor surveillance vantage point while others are taken in special cases such as in a wide-angle indoor environment where faces appear at a distance and thus the resolution is very poor.

In Dataset #2 several sequences were taken in very crowded environments such as would occur in real life video surveillance scenarios, both indoor and outdoor, where occlusions, cast shadows, and frequent changes of the head pose are very common. This dataset was collected specifically to evaluate the effectiveness of face logging systems under realistic, video surveillance scenarios. Sequences are taken with a varying number of people (from about ten to more than one thousand persons) and are composed of about 60 minutes of video of critical conditions in which there may be frequent identity mismatches due to the proximity of people.

Dataset #3 is composed of two sequences, each of about five hours, taken in an indoor environment. This dataset was collected to evaluate the long-term robustness of face logging systems.

4.3.1 Comparison with other face logging systems

We compare our face logging system to the recent approaches in [57] and [96]. The comparison is made using the publicly available Hermes head pose dataset [2] used in [96]. Following the protocol described in [96], we ran our approach on the 48 sequences, each manually annotated to indicate which

[2]The dataset is publicly available at `http://www.cvmt.dk/projects/Hermes/head-data.html`

face image was of highest quality. We then count how many times the best face image log is ranked first by our algorithm.

As can be seen in table 4.2 we outperform previous with a *correct matching rate* of 95.8%, using the quality filter $\mathbf{Q}_{pose}$ of eq. 4.8. The improvement over the use of $\mathbf{Q}_{pose}$ results from the fact that the improved quality filter is better at filtering false positives, the detectors can be run to optimize recall, and more targets are logged (see the experimental results in 6.4-C for a more detailed analysis of the precision/recall trade off). Our logger misses only two intrusions over the 48 processed sequences.

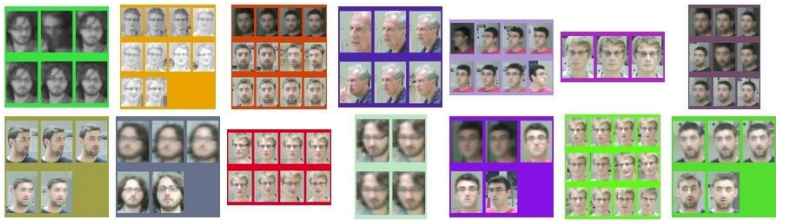
Method / Feature	Approach [57]	Approach [96]	**Our approach**
Single-Target	Yes	Yes	**Yes**
Multi-Target	Yes	No	**Yes**
Automatic Init	Yes	Yes	**Yes**
Accuracy	87.1%	88.5%	**89.6** ($\mathbf{Q}_{sym}$), **95.8%** ($\mathbf{Q}_{pose}$)

Table 4.2: Comparison with other face logging system [57, 96] on the HERMES dataset.

4.3.2 Qualitative performance analysis

In figure 7.5 we show some qualitative results. In particular, figure 7.5(a) contains some logs generated by our system on the Long #2 sequence (each color describes the log identity). We can see that logs are brief and concise, but also meaningful and accurate. Figure 7.5(b) illustrates how the face logger manages multiple target in a crowded scene. The tracked target identity is represented with a colored circle.

Finally, in figures 7.5(c) to 7.5(e) we show the summarization capability of our face logging system. We test it on the fourth person of the HERMES and we give a comparison by showing the logs generated with the two proposed quality filter measures with respect to those generated with a brute force approach that stores a face in the log for every detection associated to a track. In this case, shown in figure 7.5(c), the set of faces is overcomplete. Filtering logs on the basis of resolution and symmetry results in a more compact log, as shown in figure 7.5(d), though there is still some redundancy. Resolution and pose result, however, in the briefest logs without sacrificing accuracy, as can be observed in figure 7.5(e).



(a) Some of the logs resulting from the Long #2 sequence. Images are ordered left-to-right, top-to-bottom according to image quality measured by equation (4.8), best quality last.

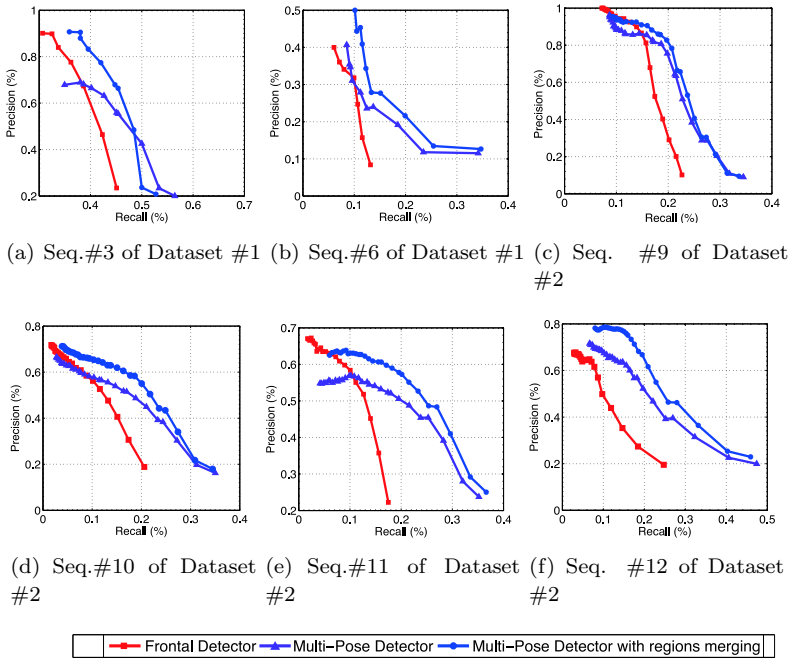
(b) Results from a crowded indoor sequence from Dataset #2.

(c) Brute force approach (159 images omitted).

(d) Resolution and symmetry

(e) Resolution and head pose

Figure 4.3: Examples of our face logging system. *(a)*: Face logs produced by our approach, each color representing a different log. *(b)*: Face pose estimation and face tracking. The white superimposed indicator shows the estimated head pose of the target. *(c)* Log generated using a brute force approach. *(d)* Log generated using resolution and symmetry. *(e)* Log generated using resolution and head pose.



(a) Seq.#3 of Dataset #1 (b) Seq.#6 of Dataset #1 (c) Seq. #9 of Dataset #2

(d) Seq.#10 of Dataset #2 (e) Seq.#11 of Dataset #2 (f) Seq. #12 of Dataset #2

Figure 4.4: Precision/recall curves, validated over 55,367 faces, comparing the frontal detector [123] with our multi-pose detector with and without detection merging as described in section 4.2.1.

4.3.3 Quantitative performance analysis

In figure 4.4 we show how our multi-pose face detector described in section 4.2.1 improves the recall and the precision against the standard AdaBoost technique of Viola and Jones. These results illustrate the utility of multi-pose detection in realistic scenarios where non-frontal faces are a common occurrence. Each face missed by a frontal detector represents a potential false negative downstream in the face logger. We evaluate detection responses over sequences #3 and #6 of the Dataset #1, and sequences #1, #2, #3, and #4 from Dataset #2. These sequences were selected because they are among the most informative ones from the dataset and together contain a total of 55,367 annotated faces. Performance is expressed using precision/recall showing the precision against recall obtained by varying the

confidence threshold of the detector. Given a face annotated with bounding box A and a detection response with bounding box D, we used the VOC-score criterion defined in equation (4.4) and we consider A as a true positive if and only if: $\text{VOCscore}(A, D) \geqslant 0.5$.

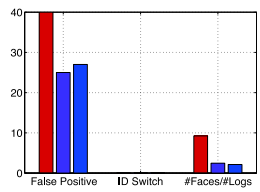
The red curves show the performance of the standard AdaBoost face detector, the violet curves the performance of our multi-pose face detector, and the blue ones the performance of our multi-pose detector with merged detections. Note that our detector not only improves overall estimation accuracy, but also improves over the recall of a frontal face detector by merging the responses from frontal and lateral cascades.

Figure 4.5(a) summarizes the face logging experiments on each of our datasets using the $\mathbf{Q}_{pose}$ quality measure of eq. 4.8. We ran our system optimizing recall (subscript r, minimizing the number of missed faces) and precision (subscript p, minimizing the number of false positives returned), reporting for each setting the following values: the number of logs produced by the face logging system (**#Logs**); the percentage of logs in which at least one identity switch occurs (**ID sw**) between two or more targets; the percentage of logs in which there is at least a false positive (**FP**) face picture; the recall of the logging system defined as the percentage of persons actually logged with respect to those in the scene (**Recall**); and a measure of log brevity with respect to the original video size ($\boldsymbol{\eta}$) expressed as the compression factor in video summarization $\eta = \text{size}_{video}/\text{size}_{logs}$. Due to the use of a particularly restrictive policy for tracker birth and death, the face logger produces a large number of tracks for the longest and most crowded sequences (Dataset #2). Although this reduces the probability of identity mismatch, it can produce multiple logs for a single actual target. However, the very low value of the identity mismatch ratio reported for each dataset – which is very low even for the most crowded scenes – is encouraging. Our system gives a low false positive rate for all sequences when configured to optimize either precision or recall. Only for the Room sequences when optimizing recall do we have a persistent false positive that initializes eleven tracks resulting in a high percentage of false positives.

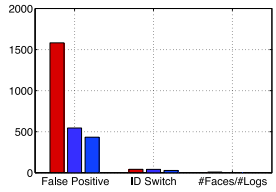
From figure 4.5(a) we see that our face logging system yields high recall on all sequences – in particular on the crowded indoor sequences despite the difficulties in the initialization of each target due to occlusion and proximity between them. The low recall on the Room sequences (when optimizing precision) is due to a combination of difficult detection conditions and the

Value ↓	**Dataset 1**		**Dataset 2**		**Dataset 3**	
Sequences →	*Door/Corridor*	*Room*	*Crowd In.*	*Crowd Out.*	*Long #1*	*Long #2*
Video Size	232MB	202MB	713MB	1.42GB	834MB	885MB
$\#\text{Logs}_p$	113	14	359	1799	153	258
$\#\text{Logs}_r$	205	35	1313	4663	268	304
FP_p (%)	0	0	0.56	0.44	2.61	0.78
FP_r (%)	3.4	31.42	20.48	7.78	18.28	7.24
ID sw_p (%)	0	0	0.56	6.22	0	0
ID sw_r (%)	0	0	2.14	7.10	0	0
Recall_p (%)	90	44.44	82.93	–	94.12	88.89
Recall_r (%)	93.75	83.25	97.8	–	94.12	88.89
η_p	92X	250X	87X	29X	99X	59X
η_r	52X	263X	44X	2X	20X	54X

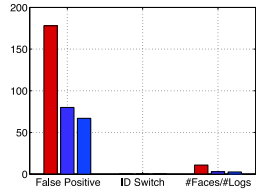
(a) Quantitative results using resolution and pose. We tested two configurations of our system: a subscript r means results are produced with the detector set to optimize recall, while subscript p means that the detector instead optimizes precision.



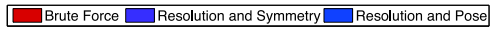
(b) Dataset #1

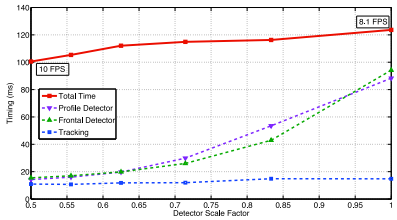


(c) Dataset #2 (Indoor)

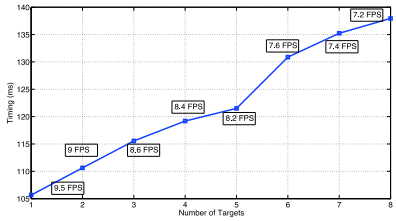


(d) Dataset #3





(f) Timing varying image resolution (scale).



(g) Timing varying the number of targets.

Figure 4.5: (a) Quantitative performance analysis of our face logging system on datasets #1, #2 and #3 using the quality filter based on pose. (b-d) Number of false positives, ID switches and faces per log with detector optimized for recall. (f) Timing for our face logger in milliseconds over a range of detector scales; dotted lines show time per stage, while the red line shows the time for the entire process. (g) Timing over a varying number of targets and the detector scale is set to 0.85.

relatively low number of individuals in the sequences. Unfortunately, calculation of recall for the crowded outdoor sequences (see figure A.2) is infeasible since it contains more than one thousand people, each appearing at several different times.

In figures 4.5(b-d) we report logging performance for the quality functions described in section 4.2.4. Histograms show the average number of false positive and ID switches over all the saved faces and the number of faces per log. To test the quality measures we ran our face logger system using the multi-pose detection set to optimize recall so that almost all persons present in the scene are logged; doing this effectively stresses the quality measure and evaluates which is the best filter of false positives and ID switches. The brute force approach provides a baseline in which a face log is created for all detections associated with a track (red bar). From this figure we see that resolution and face pose do not improve the number of faces per log over resolution and symmetry. They do, however, suppress the number of errors (false positives and ID switches) of the logging system on the indoor crowd sequences (the outdoor crowd sequences cannot be evaluated as no annotations are available). On the first and third datasets it perform similar to resolution and symmetry.

As shown in figure 4.5(f) and 4.5(g) we have also conducted an analysis of time used by our system to log faces. The evaluation was performed on an Intel Xeon Quad-core processor at 2.97 GHz, with a C++ implementation that exploits all available cores. From figure 4.5(f) we can observe that in most cases our approach runs at frame-rate of about 10 fps and that the detector timing does not significantly affect the timing of the overall system. Moreover, as shown in figure 4.5(g) there is a linear relationship between the number of targets in the scene and the computational burden.

4.4 Conclusions

In this chapter we have proposed a method for posterity logging of face imagery: a way to save a compact, semantically relevant snapshot of a scene that might be relevant for future investigations. These logs generated are *brief* and *high-quality* due to the measures of resolution, symmetry and pose used to filter face imagery. Experiments show that our multi-pose face detector is robust and, in addition to insuring high recall while maintaining respectable precision, also provides an estimate of head pose that can be

used as a facial image quality measure. Symmetry was also showed to be an effective quality measure, one that is inexpensive to compute and that can be applied in situations where head pose is not immediately available. Saving only the best face images seen of each target ensures that downstream forensic analysis will not be overwhelmed with many redundant images of the same target. The logs generated are *accurate* due to the conservative data association policy we use to associate face detections with running trackers.

The bulk of the experimental results presented in this work are on indoor sequences, in situations where face size is bounded, illumination conditions are consistent with the training images used to train AdaBoost detectors in our multi-pose face detector, and in general when face detection can be expected to be reliable. This is a common and realistic scenario, though applying our approach to scenes with greater face scale and illumination variation can be expected to increase false positive detections. Note, however, that the experimental results on outdoor sequences presented in Figure 4.5(a) are encouraging.

Chapter 5

2D/3D Florence Faces Dataset

This chapter describes a new dataset under construction at the Media Integration and Communication Center and the University of Florence. The dataset consists of high-resolution 3D scans of human faces from each subject, along with several video sequences of varying resolution and zoom level. Each subject is recorded in a controlled setting in HD video, then in a less-constrained (but still indoor) setting using a standard, PTZ surveillance camera, and finally in an unconstrained, outdoor environment with challenging conditions. In each sequence the subject is recorded at three levels of zoom. This dataset is being constructed specifically to support research on techniques that bridge the gap between 2D, appearance-based recognition techniques, and fully 3D approaches. It is designed to simulate, in a controlled fashion, realistic surveillance conditions and to probe the efficacy of exploiting 3D models in real scenarios.[1]

5.1 Introduction

Human faces are one of the most important biometrics for recognition. Face imagery is easily and non-intrusively collectible, whereas other biometrics

[1]This chapter has been published in *Joint ACM Workshop on Human Gesture and Behavior Understanding (J-HGBU'11) ACM Multimedia Workshop 2011, Arizona,USA, 2011* and in *IEEE 5th International Symposium on Communications, Control, and Signal Processing 2012, Roma, Italy, 2012.*

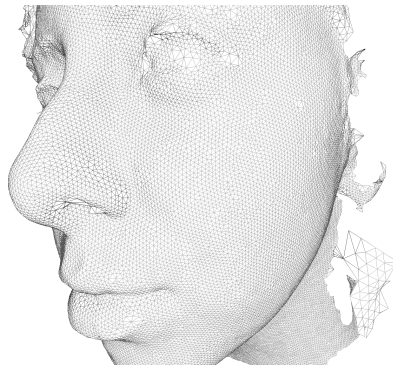


Figure 5.1: An example of the high resolution 3D models captured in the Florence Face Dataset. On average, all models have around 40,000 vertices, 80,000 facets and are accurate to 0.2mm.

such as fingerprints or iris scans are impractical to implement in many scenarios (e.g. a surveillance setting). Because of the universality of faces as a biometric, there has been a proliferation of face recognition approaches proposed in the research literature [138]. Along with the proliferation of algorithms for face recognition there has also been an explosion of datasets designed to support research in face recognition.

Face detection and recognition remains a difficult problem. Much of this difficulty is due to challenging imaging conditions and variations caused by expressions, gender and pose. More recently, 3D scanning technology has matured and the price of entry is much less. This has led to renewed interest in face recognition using 3D models of human faces. One unexplored avenue of research on facial analysis is the potential of using 3D models to augment the performance of traditional 2D, appearance-based techniques. One of are prime motivations for developing the Florence 2D/3D Face Dataset was the recent advances in tracking with PTZ cameras [44].

In this chapter we describe the Florence 2D/3D face dataset that is currently under preparation [2]. The dataset has been specifically designed to support research across a range of face analysis and recognition tasks. In the next section we survey existing face datasets and in section 5.3 we describe the structure and organization of the Florence 2D/3D Face Dataset.

[2]Visit `http://www.micc.unifi.it/datasets/3D-faces` for an interactive preview of the dataset.

Finally we conclude in section 6.5 with a discussion of the direction our work will take on this new 2D/3D face dataset.

5.2 Existing Face Datasets

In this section we briefly review some currently available face datasets. We begin first with 2D face datasets designed to support research on appearance-based facial analysis techniques.

5.2.1 2D Face Datasets

The analysis of human faces for biometric purposes is one of the oldest applications of computer vision. Given the consistently high level of interest in the topic, there is an abundance of resources for conducting research in the area. The majority of these resources are datasets consisting only of 2D images and/or video.

The CaltechFaces [126] dataset consists of 450 face images in JPEG format at a resolution of 896×592 pixels. Around 27 people are captured under different lighting conditions, with different expressions, and against different backgrounds. The Caltech face dataset is a classic example of what we refer to as a *2D face dataset*. That is, it is purely designed for conducting research on appearance-based techniques for facial analysis and recognition.

Facial analysis algorithms are difficult, in part, due to the huge variation in pose, background and illumination conditions under which faces may be captured. The Multi-PIE [63] dataset is an example of a dataset that advanced the state-of-the-art in facial recognition by make available facial images with varying factors affecting the appearance of faces in them. The original PIE database, collected at Carnegie Mellon University in 2000, was already been very influential in advancing research in face recognition across pose and illumination. Despite its success the PIE database has several shortcomings: a limited number of subjects, a single recording session and only few expressions captured. The Multi-PIE dataset, also collected at Carnegie Mellon University, contains 337 subjects, captured under 15 viewpoints and 19 illumination conditions in four recording sessions for a total of more than 750,000 images.

The FRAV2D dataset [64] contains around 100 subjects, with 32 color images at a resolution of 320×240 pixels. For each person, all photographs

are taken on the same day, although the subject is made to stand up and sit down again in order to change pose and gesture. In all cases, the background is plain and dark blue. Each image is classified into one of six groups according to the pose and lighting conditions: 12 frontal images, 4 15°-turned images, 4 30°-turned images, 4 images with gestures, 4 images with occluded face features and 4 frontal images with a change of illumination.

CASIA-FaceV5 [35] contains 2,500 color facial images of 500 subjects captured using a standard, USB webcam in one session. All face images are 16-bit color BMP images at a resolution of 640×480 pixels. This dataset is notable in that its images are acquired using a commodity webcam and as such represents a more realistic application setting than other 2D datasets in the literature.

5.2.2 3D Face Datasets

From a relatively early point in the development of facial recognition algorithms it was realized that one way to marginalize the affects of illumination, pose and imaging conditions is to record 3D models of faces instead of 2D images or video. 3D models are robust to these factors as the geometry of objects is invariant to imaging conditions. In this section we describe some existing datasets containing 3D models of faces.

The Texas 3D Face Recognition Database [66] contains 1,149 pairs of color and range images of 105 faces. The images were collected by Advanced Digital Imaging Research (ADIR), with assistance from research students and faculty at The University of Texas at Austin. The images were acquired using a stereo imaging system at a high spatial resolution of 0.32mm. The images in the dataset are of adult humans from all the major ethnic groups and genders, and each face also comes with metadata about subject gender, ethnicity, facial expression, and the locations of 25 manually located, anthropometric fiducial points.

An interesting approach to constructing a dataset of 3D face images is the Basel Face Model (BFM) [98]. The BFM is a 3D morphable face model constructed from 100 male and 100 female example faces. It consists of a generative 3D shape model covering the face surface from ear to ear and a high quality texture model. This model was designed to be used directly for 2D and 3D face recognition or to generate training and test images simulating any imaging condition. The dataset contains all training data used to learn the 3D morphable face model (the BFM) and the model fitting results for

several standard image data sets (CMU-PIE, FERET).

The Bosphorus Database [108] is a 3D face dataset that includes a rich set of expressions, variation of poses and different types of occlusions. This database is unique for a number of reasons. First, the facial expressions are composed of a subset of action units as well as the six basic emotions, and many actors are used to obtain more realistic expression data. Second, a rich set of head pose variations are represented in the dataset. Finally, different types of face occlusions are also represented. This dataset is a valuable resource for development and evaluation of algorithms under adverse conditions.

The FRAV3D [65] database contains 106 subjects acquired with a Minolta VIVID 700 scanner, which provides texture information (2D image) and a VRML file (3D image). It is a multimodal database containing 2D, 2.5D and 3D captures of each subject. A total of 16 captures per person were taken in every session, with different poses and lighting conditions One of the main advantages of this database, compared to others, is the extent to which different poses are captured in the 16 scans taken of each subject.

GavabDB [95] is a 3D face database containing 549 three-dimensional images of faces. The dataset contains images of 61 different individuals (45 male and 16 female), with 9 images per subject. Each 3D model is mesh of connected 3D points of the facial surface without texture. The database provides systematic variations with respect to the pose and the facial expression. In particular, the 9 images corresponding to each individual are the following: 2 frontal views with neutral expression, 2 x-rotated views (30°, looking up and looking down respectively) with neutral expression, 2 y-rotated views (90°, left and right profiles respectively) with neutral expression and 3 frontal gesture images (laugh, smile and a random gesture chosen by the user, respectively).

The Casia3D [35] dataset consists of 4624 scans of 123 persons using the Minolta Vivid 910 scanner: the dataset contains multiple pose variations, expressions and illuminations, as well combinations of expressions under illumination and poses under expressions. From each scan, one 2D color image and one 3D facial triangulated surface are generated.

5.2.3 Hybrid 2D/3D Datasets

Finally, in the recent literature on facial analysis datasets there have begun to appear a number of “hybrid” datasets that, to one degree or another, mix

both 2D and 3D information about each subject. An example is the Face Recognition Grand Challenge (FRGC) 2005 [99] dataset, which is to our best knowledge the most complete hybrid face dataset. In the FRGC dataset a subject session consists of four controlled still images, two uncontrolled still images and one three-dimensional model.

Hybrid datasets are used to determine whether 3D imagery is more effective then 2D imagery for facial analysis, and more generally if 3D information can be used as information complementary to 2D images in order to boost performance. The Florence Face Dataset is also a hybrid 2D/3D dataset, and in the next section we describe how we have designed it specifically to explore these questions in realistic scenarios.

5.3 Design of the Dataset

When we use the term "2D" we refer to video or image sequences of database subjects. Our dataset is designed with two main goals in mind. First, we would like to make available accurate and complete 3D models of faces to researchers who are primarily interested in the analysis of 3D meshes and textures of human faces. That is, our dataset is designed to be useful for research on pure 3D analysis techniques.

Second, however, we have designed our dataset as something that goes beyond the scope of 3D analysis techniques, allowing researchers to investigate the possibility of reducing the gap between 2D computer vision algorithms and those methods that work on precise 3D models. In particular, our dataset is thought of in the context of evaluating the use of 3D information in computer vision problem like 3D face pose estimation [30,51] and 3D face recognition [23], but *directly from video data or still images.*

To this end, the pipeline of data acquisition is designed to provide both 3D data and 2D videos consistent with each others: as shown in figure 5.2. First a 3D model is captured of the subject using a 3D scanner. Second, we record HD video of the subject as he simulates specific head rotations (this corresponds to a cooperative environment). Four levels of zoom are used in order to capture the subject face at multiple image resolutions. After this, the subject is then recorded from two PTZ Cameras, one indoor and the other outdoor. These two scenarios represent a more non-cooperative subject and he is asked to be spontaneous. Three levels of zoom are captured in each video in order to cover a broad range of face resolutions.

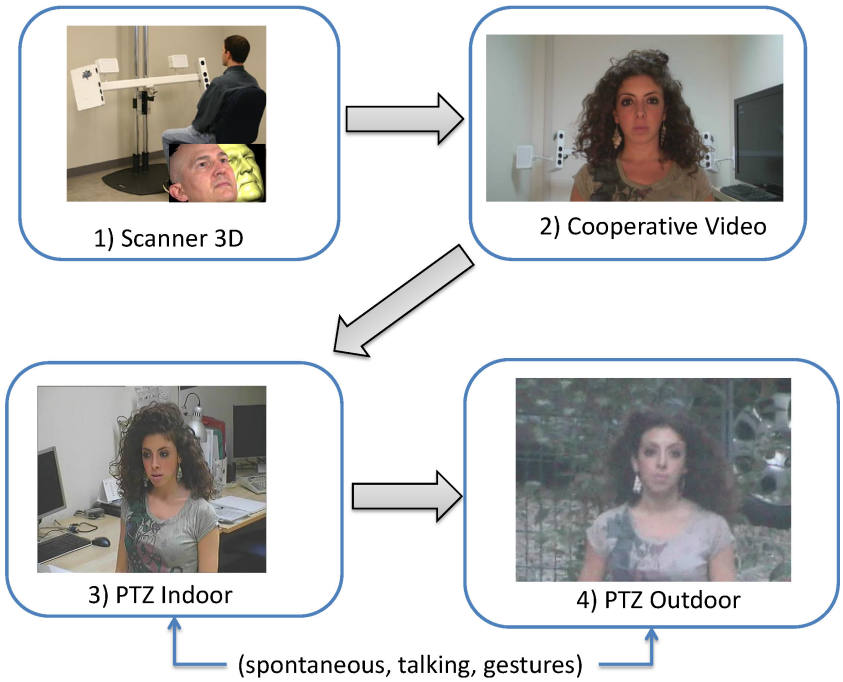


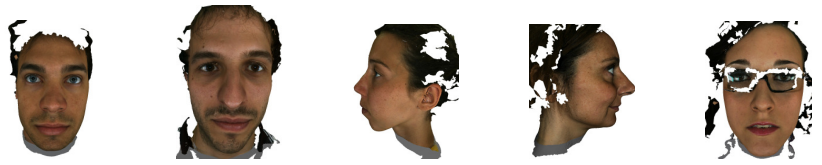
Figure 5.2: The acquisition process.

The Florence 2D/3D Face Dataset is currently a continuous work in progress and consists of 53 adult subjects as of January, 2012. The aim is to complete the dataset with a total of 100+ subjects. The current subjects are Caucasian individuals with ages ranging from 22 to 60 years. We are attempting to make the dataset balanced in gender, though currently we have 75% male and 25% female subjects. The data collected from each subject is a number of 3D face models and 2D face imagery, each of which is described in the following sections.

5.3.1 3D Face Models

In terms of 3D data, our dataset contains the following high-resolution 3D models (both meshes and textures). Some example models are shown in figure 5.3. The models are summarized below:

- Two frontal models with the head in approximately the same pose. This



(a) Frontal 1 (b) Frontal 2 (c) Left Side (d) Right Side (e) Glasses

Figure 5.3: Examples of each type of 3D model for four random subjects.

is designed so that one may be used for training and the other for testing. In figure 5.3(a) and figure 5.3(b) the frontal models are displayed for a variety of subjects.

- One left side model, which covers approximately the whole side of the faces: from mid-nose to the back of the head. See figure 5.3(c) for examples.
- One right side model, with the same properties as above. See figure 5.3(d) for examples.
- If the subject wears glasses, we also capture a 3D model with glasses (see figure 5.3(e)).

All 3D models are provided in three different formats: OBJ, PLY and VRML. They contain the mesh and the high resolution texture.

To capture each 3D model we used a state-of-the-art 3dMD Scanner [2]. It is not a laser rangefinder-based scanner, but instead works with the principle of structured light in a way similar to Kinect. It superimposes a predefined pattern using a projector in the visible spectrum and then uses two pairs of stereo cameras to perform the 3D reconstruction of the facial surface. The texture is acquired with a stereo RGB camera. The device can cover 180-degrees of a frontal face, capturing from ear-to-ear at a speed of about 1.5 milliseconds at the highest resolution of 4Mpixel.

The geometry of the mesh is extremely accurate: on average about 0.2mm of RMS error in the reconstruction or better, depending on the exact pre-calibration and configuration. Considering this, our 3D models are very accurate and precise in the number, yielding on average a mesh of about 40,000 vertices about 80,000 facets. The texture acquired is a stereo image with a resolution of 3341×2027 pixels.

Dataset	Availability	# Subjects	Image/Video	3D Model	HD Image or Video	Annotations
Florence 2D/3D	**On Written Request**	53	No/**Yes**	**Mesh+Texture**	**Yes**	None
Texas 3DFRD [66]	On Written Request	118	No/No	Range+Texture	No	Fiducial Points
FRAV3D [65]	Not Publicly Available	106	No/No	Range+Mesh+Tex.	No	None
FRGC 2005 [99]	On Request	466	Yes/No	Range+Texture	Yes	None
GavabDB [95]	Public	61	No/No	Mesh Only	No	None
Casia3D [35]	On Registration	123	No/No	Mesh+Texture	No	None
Bosphorus [108]	Public	105	No/No	3D Map+Texture	No	Various
BFM [98]	On Written Request	200	No/No	Mesh+Rendering	No	Various

Table 5.1: A comparative examination of available datasets supporting research on human faces, especially considering those who provide 2D appearance information like video or images along with 3D data.

5.3.2 2D Face Imagery

Once that the 3D models are acquired, the face of the subject is recorded in several 2D videos in order to obtain the same face as seen from a camera. In particular, we record three different videos in order to capture three different imaging scenarios. The video sequences are designed to provide video across a range of standard camera qualities and resolutions. They are also recorded at multiple levels of zoom in order to investigate the 2D face resolution necessary to support accurate facial analysis. Finally, the imaging scenarios used are designed to simulate and indoor environment with a cooperative subject, and indoor environment with an uncooperative subject, and an outdoor environment with an uncooperative subject.

The three type of videos recorded for each subject are defined as follows:

- One indoor HD video (1280×720 pixel resolution) of cooperative subject recorded at 4 levels of zoom. To generate a uniform source of light without highlights, the subject is irradiated with two lamps in front of her. The subject is asked to generate some out-of-plane head rotations, viewing six points: top-right, top-left, middle-right, middle-left, bottom-right, bottom-left. The frame rate for this video is about 20 fps and is acquired with a AXIS Q1755 HD camera. Example frames from some of these sequences are shown in figure 5.4.
- One indoor video (704×576 - 4CIF resolution) of a uncooperative subject from a PTZ camera using 3 levels of zoom. Here the subject is asked to be spontaneous. The frame rate for this video is about 20 fps and is acquired with a AXIS PTZ Q6032-E (see figure 5.5 for examples).
- One Outdoor Video (736×544 pixel resolution) of uncooperative subject from a PTZ camera using 3 levels of zoom. Here the subject is again asked to be spontaneous, but this time the recorded video is very chal-

lenging because of the uncontrolled lighting conditions and the presence of shadows and highlights. The frame rate for this video is about 5-7 fps and it is acquired with a SONY RZ30-P camera. Examples are given in figure 5.5.

All videos are encoded in the MJPEG format in order to maintain the original definition of the source, paying some complexity in the storage. The uncooperative videos indoor and outdoor are recorded with PTZ camera mainly because of the proliferation of this type of device and because of its capability to redirect the field of view, zooming-in on the face and achieving high resolution images at a distance.

5.3.3 Comparison with Other Datasets

Building an unbiased dataset is not a simple task even though the work on faces is one of the earliest problem addressed in computer vision. Considering this, we would like to remark that almost all datasets are built considering a particular goal to achieve.

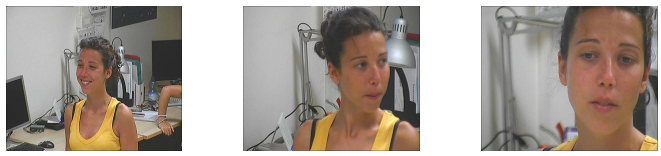
In this section we point out the basic differences of our dataset compared to the the others presented in section 7.2 Our dataset is intended for research on 3D and 2D human face processing tasks including face recognition under adverse and good conditions, 3D face reconstruction from videos/images, and face pose estimation. The Florence 2D/3D Hybrid Face Dataset is unique in several aspects. First, we provide multiple, high-quality 3D models, both meshes and textures, plus multiple high-resolution videos of each subject. Also, the image sequences are designed to mimic a range of difficult, real-world application scenarios involving facial analysis. Other substantive differences are summarized in Table 5.1.

5.4 Discussion

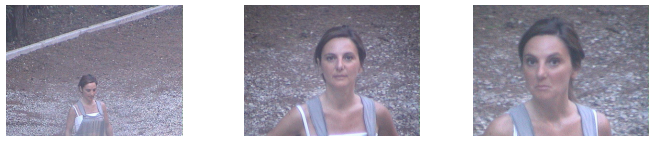
In this chapter we discuss and review 2D and 3D face datasets available in literature. By far the most common type of dataset available to support research on facial analysis algorithms are standard, 2D datasets consisting of multiple, static images of each subject. Pure 3D datasets for facial analysis are starting to appear, though an unexplored area of research remains the potential intersection of 2D and 3D facial analysis algorithms. This is the niche in which we have positioned the Florence 2D/3D Face Dataset. The

(a) Zoom level 1 (b) Zoom level 2 (c) Zoom level 3 (d) Zoom level 4

Figure 5.4: Samples of the high definition video taken of subjects in a cooperative setting.



(a) Zoom level 1 (b) Zoom level 2 (c) Zoom level 3



(d) Zoom level 1 (e) Zoom level 2 (f) Zoom level 3

Figure 5.5: Examples of video sequence of a subject taken from a PTZ camera in an indoor environment (First row) and outdoor (Second row). Three different levels of zoom are used to generate views of the subject face at different resolutions. Subject behavior is natural and unscripted.

Florence dataset is, at the time of writing, still a work in progress, but is scheduled to be a complete resource of 100+ subjects. Our hypothesis is that 3D information in the form of high-resolution models of subjects can be useful for a range of facial analysis techniques that are classically very challenging using only traditional 2D imagery.

Chapter 6

Matching Face Imagery with 3D Textured Models

In this chapter we consider the problem of face recognition in imagery captured in uncooperative environments using PTZ cameras. For each subject enrolled in the gallery, we acquire a high-resolution 3D model from which we generate a series of rendered face images of varying viewpoint. The result of regularly sampling face pose for all subjects is a redundant basis that over represents each target. To recognize an unknown probe image, we perform a sparse reconstruction of SIFT features extracted from the probe using a basis of SIFT features from the gallery. While directly collecting images over varying pose for all enrolled subjects is prohibitive at enrollment, the use of high speed, 3D acquisition systems allows our face recognition system to quickly acquire a single model, and generate synthetic views offline. Finally we show, using two publicly available datasets, how our approach performs when using rendered gallery images to recognize 2D rendered probe images and 2D probe images acquired using PTZ cameras. [1]

[1]This chapter has been published as "Using 3D Models to Recognize 2D Faces in the Wild" in *Proc. of CVPR Int'l Workshop on Socially Intelligent Surveillance and Monitoring, Portland, USA, 2013.*

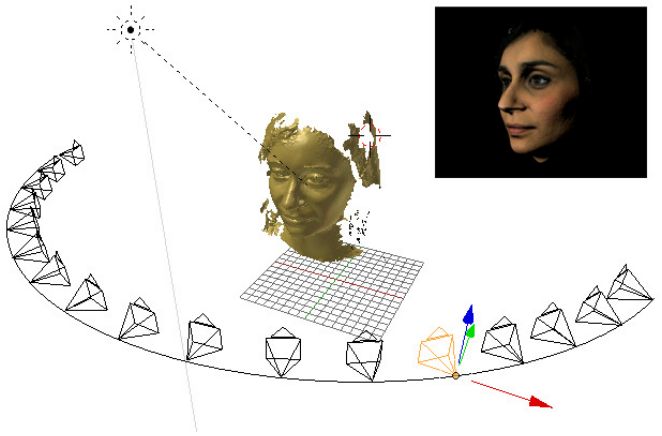


Figure 6.1: Synthetic data generation process: given a 3D model, we sample the yaw angle by rendering 25 poses. The highlighted camera gives the facial image shown in top right corner.

6.1 Introduction

Automatic face recognition is one of the classic, fundamental problems in the computer vision community. In recent years even more effort has gone into studying techniques and systems for accurately modeling facial appearance and for recognizing faces in diverse environments [100]. A general statement of the automatic face recognition problem, from a computer vision standpoint, can be formulated as follows: given a *probe image or video* of a scene, verify the identity of one or more of the persons in it using stored *gallery of known individuals*. Despite its long history as a central problem in computer vision, face recognition remains a subject of great practical and theoretical interest [100].

The basic process of face recognition consists of:

- **Enrollment** of individuals in the gallery of known people. Enrollment usually takes the form of the capture of a sequence of high resolution images of each person, or a 3D model of each face if the system is designed for recognition of 3D probe images. A critical point for applicability of face recognition systems in practice is that enrollment be as efficient as possible.

- **Learning** of discriminative or generative models of gallery subjects to be used for later recognition of faces in probe images. A variety of methods can be used for this stage, and in case of 3D face recognition the learning process often involves the estimation of an average 3D face that will be used to register probe image faces at recognition time.
- **Recognition** of unknown individuals in probe images. In this phase unknown faces in probe images are classified using models learned on the gallery image set. There are also myriad recognition scenarios, though they can be coarsely categorized into cooperative and uncooperative scenarios. In *cooperative* scenarios the unknown person is assumed to actively submit to facial image capture at recognition time and the resulting probe images are usually frontal and of very high quality. In *uncooperative* scenarios, recognition is passive and probe images must be captured using passive sensors in the environment. As with enrollment, it is important that recognition be as efficient and non-intrusive as possible.

In this chapter we take a hybrid approach that exploits 3D face models to recognize faces in PTZ camera imagery. From a high resolution 3D model of faces, we artificially generate multiple views of each subject by rendering the enrolled 3D models from varying viewpoints. The acquisition process for rendered 2D model views is illustrated in Fig. 6.1. From these rendered face images we extract SIFT descriptors at salient image positions, and, rather than quantizing these descriptors against a visual vocabulary, we then represent each individual as an unordered bag of SIFT features. By varying the viewpoint of subjects in the gallery, we reduce the need for frontal face imagery for use as probes. Probe images are also represented as unordered bags of SIFT features, and recognition is performed through sparse reconstruction of probe image features from gallery image features. The use of sparse reconstruction allows our approach to leverage the multiple views of each subject in the reconstruction of unknown probe images.

In the next Section we discuss work related to face recognition and sparse discriminative classifiers. In Section 6.3 we describe how we acquire high-resolution 3D models of gallery subjects, generate rendered images from multiple viewpoints of each, and finally classify unknown probe images using these rendered views. We describe a series of experiments performed on two face datasets in Section 6.4, and finally conclude with a discussion of ongoing work in Section 6.5.

6.2 Related Work

In this section we briefly review the literature on hybrid recognition approaches, by which we mean automatic recognition systems using both 3D and 2D face data. For a more thorough survey of face recognition in general, the interested reader should consult the excellent reviews in [3, 137].

The method in [28] estimates 3D shape and texture of faces from single images. Rather than directly acquiring a 3D model from faces at enrollment, an estimate of a 3D face model is computed by fitting a morphable 3D model, learned from a set of textured 3D scans of faces, to images. Recognition is performed by matching the shape and texture information after fitting the 2D probe images to the 3D model.

In [74] the authors propose a method for view and pose invariant face recognition that combines component-based recognition and 3D morphable models. The approach first uses a 3D morphable model to generate 3D face models from only two input images of each person enrolled in the gallery database. By rendering the 3D models under varying pose and illumination conditions they create a large number of synthetic face images which are used to train a component-based face recognition system. Differently from our approach they generate a coarse 3D model from two 2D views of face and perform a two stage classification in which they first individuate the face component in the test image using an SVM classifier then detect the configuration of components to feed a geometric classifier.

The authors of [33] propose a face recognition solution combining both 2D and 3D face data. They develop a PCA-based approach tuned separately for 2D and for 3D. A multi-modal decision is obtained by first matching a 2D probe against the 2D gallery, and then the 3D probe against the 3D gallery. A confidence is computed for the 2D and 3D recognition scores and these confidences are used as weights in the sum of distances to obtain final classification score. Unlike our approach they use both 3D and 2D images in both the probe and gallery sets and only use the texture information of the 3D model as 2D views.

In [93] the authors propose a method to learn a person detector from synthetic data generated from virtual scenarios. More specifically, they record training sequences in virtual scenarios to learn an appearance-based pedestrian classifiers based on HOG and linear SVM. By testing the learned model on images containing real pedestrians they demonstrate that is possible to learn a model for detection also from synthetic data. One of the objectives

of our work is to extend this approach from detection to recognition tasks by generating synthetic views from high resolution 3D models of faces.

The ℓ_1-regularized sparse basis expansion has been used in literature to perform person recognition on well-cropped 2D face images coming from the same source. In particular, Wright *et al.* [128] show how sparse representation can be used as a powerful classification tool for face recognition. This approach has been extended several times, integrating correntropy [69] and kernel-based sparse reconstruction [78]. Elhamifar and Vidal [50] extend the Sparse Discriminative Classifier of [128] by constraining the method to find a representation of a test example using the minimum number of blocks from the dictionary (each block corresponds to multiple instances of the same subject).

6.3 2D Face Recognition from 3D Models

In this Section we describe our approach to hybrid 2D/3D face recognition. The first step in our approach is the acquisition of high resolution 3D models of each individual enrolled in the gallery, and then the synthetic generation of multiple 2D views of each individual. The final step is face recognition using the synthetic redundant basis to identify the probe.

6.3.1 2D Face Synthesis and Feature Extraction

A high resolution 3D model for each individual is quickly acquired at enrollment using a 3D scanner. From each model we artificially generate n synthetic images across varying viewpoints. These images of the i-th person in the gallery are:

$$\mathcal{I}_i = \left\{ u_i^1, u_i^2, \ldots, u_i^n \right\}, \text{ for } i \in [1, \ldots, P]\,. \tag{6.1}$$

In principle, the rendered images of each subject can be generated by varying both the yaw and pitch of each 3D model, and also by varying the illumination direction and illuminant. In this work, however, we consider only varying yaw angle for generating synthetic 2D images of each subject. We generate views by uniformly sampling 25 yaw angles in the range $[-90 + 90$. This process is illustrated in Fig. 6.1.

The final representation of individuals is an unordered bag of SIFT descriptors calculated at salient image points identified using a Harris-Laplace

Figure 6.2: 2D face views synthesized from the 3D model. Images are generated by varying the yaw angle of the 3D model rendering a 2D image.

corner detector. The Bag of Features corresponding to the i-th person in the gallery:

$$\mathcal{X}_i = \left\{ \mathbf{x}_i^1, \mathbf{x}_i^2, \ldots, \mathbf{x}_i^{n_i} \right\}, \text{ for } i \in [1, \ldots, P],$$

where each $\mathbf{x}_i^j$ is the j-th SIFT descriptor extracted from the images of the i-th gallery individual. To simplify notation we do not use an index on SIFT features to indicate from which image $\mathcal{X}$ comes.

In Fig. 6.2 we illustrate some of the rendered images derived from a model in the Florence 2D/3D face dataset. Note the high quality of the resulting images, which is due to the very high resolution of the models in the dataset (each model has around 70,000 facets, and a 4MPixel texture, on average).

Figure 6.3: Some of the 2D face views obtained from a PTZ camera at different level of zoom.

Feature extraction from probe images is performed in a similar fashion, though of course without the synthesis process from 3D models. Assume we have a probe image that contains a face region corresponding to a single individual. We use the Viola-Jones face detector [124] to identify frontal and profile faces [10]; then we extract SIFT descriptors at salient points identified with the Harris-Laplace corner detector in the detected face region. The probe image is represented as a bag of SIFT features:

$$\mathcal{Z} = \{ \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m \}, \tag{6.2}$$

where $\mathbf{z}_j$ is the j-th SIFT descriptor extracted from the probe image.

6.3.2 Face Recognition by Sparse Reconstruction

Given the gallery representation as bags of unordered SIFT features $\mathcal{X}_i$ and a probe image $\mathcal{Z}$, also represented as a bag of SIFT features, we perform face recognition using a sparse discriminative classifier, similar to that of [128]. We start by computing a ℓ_1-regularized sparse basis expansion of each probe SIFT $\mathbf{z}_i$ as a sparse linear combination of SIFT descriptors in $\mathcal{X}$:

$$\widehat{\mathbf{A}} = \arg\min_{\mathbf{A}} ||\mathbf{Z} - \mathbf{X}\mathbf{A}||_2 + \lambda||\mathbf{A}||_1, \tag{6.3}$$

where $\mathbf{X}$ is a column-wise concatenation of all gallery SIFT features from all $\mathbf{X}_i$, and thus is a matrix of size $128 \times \sum_{i=1}^{P} n_i$, and $\mathbf{Z}$ is similarly a column-wise concatenation of the m SIFT points from the probe image (and thus a matrix of size $128 \times m$). Despite the potentially large number of SIFT points (especially in the gallery), there exist very efficient techniques for solving these types of ℓ_1-regularized reconstruction problems [92]. We discovered a good value for λ to be 0.1 and fixed this value for all the experiments.

To perform classification, we examine the reconstruction error obtained by limiting the basis expansion to SIFT points extracted from gallery images corresponding to a single individual:

$$\varepsilon_i = ||\mathbf{Z} - \mathbf{X}\mathbf{I}_i\widehat{\mathbf{A}}||_2, \text{ for } i \in \{1, \ldots, P\}, \tag{6.4}$$

where $\mathbf{I}_i$ is a diagonal matrix with ones on the diagonal corresponding to SIFT descriptors in $\mathbf{X}$ extracted from images of subject i, and zeros everywhere else. This matrix effectively selects only those coefficients in the solution matrix $\widehat{\mathbf{A}}$ that correspond to the i-th person in the gallery. The identity of the probe image is classified as the one yielding the lowest overall error ε_i.

If we have multiple probe images of each subject, we apply the method described above for each image and accumulate the reconstruction errors across all probe images. Then we assign the identity to the person by taking the minimum of the ratio between the probe image yielding minimum reconstruction error and the probe image yielding the second best reconstruction error.

6.4 Experimental Results

In this section we report on a variety of experiments we performed on two face datasets. For each experiment we define the number of tested images

per subject as N, while the number of images per subject in the gallery is M. We evaluate our approach using two test modalities:

- *Single image vs Multi image*: considering each single image in the probe tested independently ($N = 1$), and having multiple images per subject in the gallery ($M > 1$).
- *Multi image vs Multi image*: using multiple images in the probe ($N > 1$) in addition to multiple gallery images ($M > 1$), modeling scenarios in which multiple face images of the same subject can be reliably associated.

We express the performance figures of our approach in term of ROC (Receiver Operating Characteristic) curves and by reporting the Recognition Rate at First Rank.

6.4.1 Experiments on 2D Images

The first set of experiments we performed was on the FacePix dataset of 2D face images [88]. This dataset is particularly appropriate for testing the central idea of our approach since each subject has been directly imaged under a variety of poses and illumination conditions. In particular, the FacePix dataset provides facial poses for each subject from +90o -90t increments of one degree. This results in 181 images per subject considering only pose variations.

In this experiment both the gallery and the probe sets contain real 2D images from the same dataset. The objective of this experiment is to show the ability of our approach to scale with respect to the number of images present in the gallery and to validate our belief that a redundant gallery can provide excellent recognition performance. We perform 2-fold cross validation considering all the images per subject where the pose ranges from -90o +90 After selecting these poses, we vary the number of images in the gallery by sampling the pose.

The results of our approach are shown in Fig. 6.4. The ROC curves represent the improvement in performance over varying numbers of images in the gallery $M = \{3, 6, 12\}$. The probe images are tested independently of each other using the *Single image vs Multi image* modality, and thus $N = 1$. Considering the number of images in the gallery, our method achieves recognition rates at first rank of 75.9% with $M = 3$, 92.2% with $M = 6$ and 98.5% with $M = 12$. These results indicate, as expected, that given enough variety in samples of each individual, high classification accuracy can be achieved.

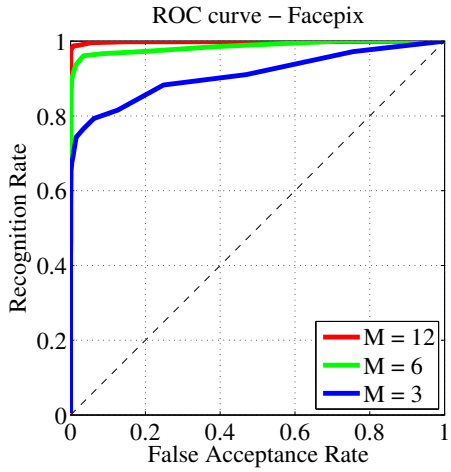


Figure 6.4: ROC curves for face recognition on FacePix. The different curves represent different numbers of gallery images per subject (M).

6.4.2 Experiments on Rendered 2D Images

For these experiments we use the 3D models from the Florence 2D/3D Face Dataset [11]. The models in the database are raw 3D meshes along with associated textures.

In order to asses the potential of our approach, we duplicate the FacePix experimental scenario with face imagery renderd using the 3D models from the Florence dataset. We rendered images from 22 of the subjects of this dataset using the approach described in Section 6.3.1. Sampling 25 yaw angles per subject, we obtain a gallery of 550 images. On this dataset we perform again 2-fold cross validation by varying the number of images per subject in the gallery in the range $M = \{2, 3, 13\}$. The rendered images are very similar to each other and face recognition performance saturates quickly. We achieve excellent recognition accuracy when considering half of the images ($M = 13$) in the gallery set per subject. The probe images are tested independently of each other in the *Single image vs Multi images* modality, and thus $N = 1$. The ROC curves for these experiments are given in Fig. 6.5(a). Varying the number of images in the gallery, we obtain recognition rates of 66.6% with $M = 2$, 84.0% with $M = 3$ and 100% with $M = 13$.

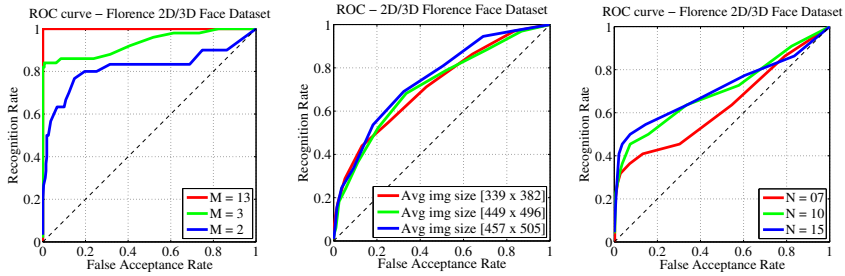


Figure 6.5: ROC curves showing performance on the Florence 2D/3D Face Dataset: each curve represents a result in function of (a) number of images (M) per subject in the gallery when recognizing 2D rendered images (b) the image size when recognizing face imagery from a PTZ camera (c) and of the number of images (N) present in the probe when recognizing face imagery from a PTZ camera

6.4.3 Rendered 2D Gallery versus 2D Probes

In this Section we report preliminary experimental results on face recognition in a video streams from the Florence 2D/3D Face Dataset captured from a PTZ camera viewing one person, as shown in Fig. 6.3. Recognition is performed using a gallery of rendered images from 3D models as described in Section 6.3.1. This scenario is very challenging considering that subjects were told to act naturally and we are basically comparing multi-modal data: probes imaged by the PTZ camera, and gallery images rendered using 3D models.

In these experiments, we tried using both a single probe image for test ($N = 1$), and multiple probe images ($N > 1$). In all these experiments we used the synthetic rendered images as described in Section 6.3.1 as gallery, thus each subject has 25 images rendered across with varying yaw as shown in Fig. 6.2.

Single image vs Multi image. The performance using single probe image per subject is shown in Fig. 6.5(b). In this experiment we also attempted to quantify how the system performs under zoom variation (given that zoom variations affect the imaged face size) by sampling faces uniformly across the entire PTZ sequence and hence including testing probe images at difference

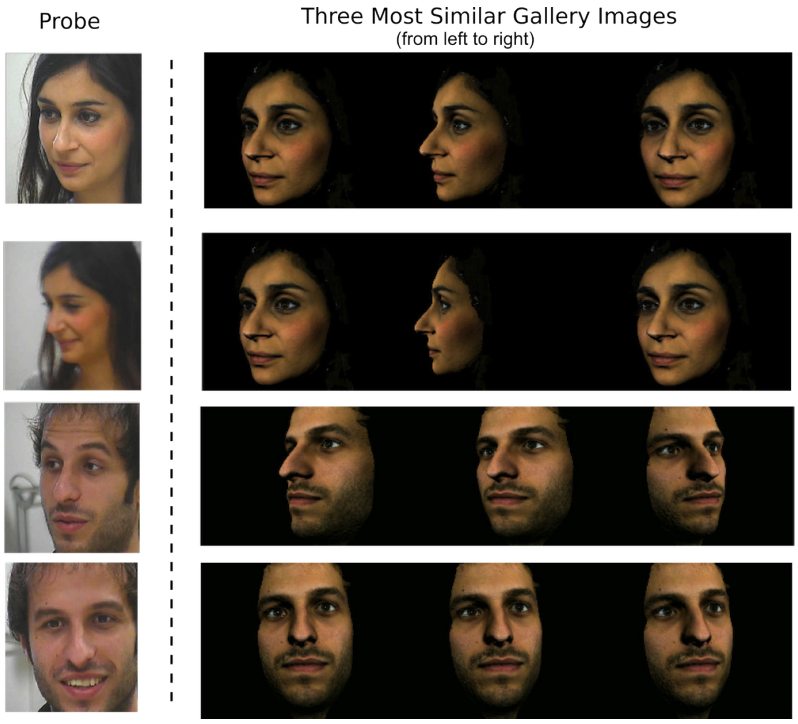


Figure 6.6: Face recognition results. *Left:* the probe image to identify. *Right:* the most similar images in the gallery (from left to right) in terms of the coefficient energy used the reconstruction ($\mathbf{I}_i\widehat{\mathbf{A}}$). Note that the face pose of the image with highest coefficient energy tends to be very similar to the pose of the subject in the 2D image.

zoom levels. In the legend of Fig. 6.5(b) we report the average size of the faces in probe images. In these ROC curves, there is little difference between the three sets of zoom levels, each achieving a recognition rate between 20% and 28%. This is likely due to the fact that other factors, such as facial expression and extreme pose variation, affect accuracy more than variations in face size.

In Fig. 6.6 we show four cases of true positive along with the three most similar images from the gallery from left to right. Note the probe images are captured "in the wild" with expressions, large pose variations and motion blur. It is interesting to note that most similar face image in the probe

usually has a face pose similar to that of the face in the image.

Multi image vs Multi image. In this experiment we evaluate the performance of our approach using multiple images in the probe. This assumption of multiple images in the probe is well known in literature for person re-identification [20] and it seems also a reasonable assumption in real world scenarios if we consider a tracker that can track and schedule a PTZ camera to follow the target face [10].

In Fig. 6.5(c) we report the performance of our approach over varying number of images used in the probe ($N > 1$). From this figure we see that using more than one image to describe an unknown person improves overall accuracy. In particular just considering $N = 7$ we outperform the *Single image vs Multi image* approach, with a recognition rate of **31.8**%. If we continue to add images from the video stream, the chance of get the right person goes up to **36.3**% with $N = 10$ and to **45.5**% with $N = 15$.

6.5 Conclusions and Future Work

In this chapter we described a hybrid approach to face recognition that uses rendered images of 3D models to form a gallery of images with varying pose for each enrolled subject. SIFT feature descriptors are extracted from these images and form a bag of features representing each gallery image. Probe images are similarly represented as unordered bags of SIFT descriptors. An ℓ_1-regularized reconstruction of probe image descriptors is used to derive a sparse discriminative classifier that effectively incorporates the information present in multiple views into the recognition process. An advantage of our approach is that no discriminative model is learned and adding new subjects to the gallery requires only concatenation of SIFT features to the existing gallery.

Experiments on a standard 2D face dataset demonstrate that our approach is very effective when very many views of each subject are incorporated into the gallery, and similar experiments on rendered 2D images for both gallery and probe show that the approach generalizes to synthetic imagery as well. Experiments on recognizing real 2D face imagery using rendered gallery images show promising results, particularly when incorporating multiple probe images per subject.

Our ongoing work is related to determining the best face images to extract

from PTZ sequences and quantifying more conclusively how performance is affected by varying face resolution and quality. We are also looking at better ways of structuring SIFT descriptors in the gallery (for example according to pose) and of structuring sparse solutions in discriminative ways (for example using the group lasso).

Chapter 7

3D Assisted Pose Independent Face Recognition

In this chapter we address the problem of pose independent face recognition with a gallery set containing one frontal face image per enrolled subject while the probe set is composed by just a face image undergoing pose variations. The approach uses a set of aligned 3D models to learn deformation components using a 3D Morphable Model (3DMM). This further allows fitting a 3DMM efficiently on an image using a Ridge regression solution, regularized on the face space estimated via PCA. Then the approach describes each profile face by computing LBP histograms localized on each deformed vertex, projected on a rendered frontal view. In the experimental result we evaluate the proposed method on the CMU Multi-PIE to asses face recognition algorithm across pose. We show how our process leads to higher performance than regular baselines reporting high recognition rate considering a range of facial poses in the probe set, up to $\pm 45^\circ$. Finally we remark that our approach can handle continuous pose variations and it is comparable with recent state-of-the-art approaches. [1]

[1]This chapter has been submitted to the *International Conference on Pattern Recognition 2014, Stockholm, Sweden*, as "Pose Independent Face Recognition by Localizing Local Binary Patterns via Deformation Components".

7.1 Introduction

Face recognition has been considered a key problem in computer vision for decades. Even if frontal face recognition seems a issue nearly solved if addressed in constrained conditions, the general problem is still open for faces captured in the wild. A "face in the wild" typically means that the subject is captured under challenging conditions such as aging, pose, expression and illumination variations. Considering these challenges, the one that mostly affects recognition performance is pose variation. In fact it is demonstrated [84] that, when the face is in a non-frontal view, face recognition performance drops drastically because discriminative descriptors, such as Local Binary Pattern (LBP) and Gabor filters, suffer from misalignment issues. In addition to these, face recognition across pose also leads to another subtle problem which is the ambiguity of landmarks caused by the self-occlusion of the face: when the face assumes a profile pose, landmark detectors respond with the same number of landmarks with respect to the ones detected on a frontal face, but with different semantic meaning (if the right part of the jaw is occluded, the detector will return a landmark on a cheek instead of on the jaw). This problem also becomes harder if we consider that in the gallery we have just one sample to describe each subject.

In this chapter we address the problem of pose invariant face recognition with a gallery set containing *one* frontal face image per enrolled subject, while the probe set is composed by just a face image undergoing pose variations. This scenario, defined as is, is an ill-posed problem considering the gap between the kind of information present in the gallery and the one available in the probe. Considering these issues, the main contributions are the following:

- we are the first to propose a way to automatically label and semantically align a set of 3D face models without any manual session. This solves the landmark ambiguity reported previously [7, 134].
- We propose a novel strategy to fit a non-rigid transformation on a face image which exploits the face modeling power of the proposed 3D Morphable Model (3DMM).
- Similarly to [134], the 3DMM is efficiently fitted on a image using a Ridge Regression solution, that globally preserves the face shape while locally minimizing the landmark reprojection error.
- By exploiting the previous contribution, instead of computing LBP on a uniform grid [4], we localize the LBP histograms on the deformed

vertices. This gives more precision to the method, obtaining features vectors of the same dimension irrespective of the image size.

The chapter is organized as follows: in Sect. 7.2 we review the most recent papers about face recognition across pose, in Sect. 7.3 we describe the method to process a 3D face database to get a set of semantically aligned faces with automatic labeled landmarks. In Sect. 7.4 we describe our approach to learn a 3DMM. Then in Sect. 7.5 we address the problem of fitting this model by minimizing the reprojection error on detected landmarks, while preserving the face shape. Once the model is fit on a generic non frontal image, in Sect. 7.6 we design our face recognition scheme by computing LBP histograms on the deformed vertices. Finally in Sect. 7.7 we perform extensive evaluation experiments respect to regular baselines such as rectification with 2D similarity, average model and other recents approaches.

7.2 Related work

Usually face recognition performance is satisfactory for near frontal faces [128] but drops drastically when the face is not showing a pose similar to the one in the gallery. Currently computer vision community has proposed several innovative methods to recognize faces across pose.
Authors in [6] propose to extend the patch-based approach of Kanade and Yamada [77] by adding a data-driven extension in which it is not only modeled how a face patch varies in appearance, but also how it is deformed geometrically as the viewpoint varies. In their case the deformation is encoded locally in an affine matrix that warps the patch without preserving the face integrity.
Instead of just using a local affine warp, the authors in [84] are the first to introduce a 3D generic face model to improve the patch-based alignment problem, rather than relying on 3D cylindrical or 3D ellipsoid model. They compare faces in different viewpoints using a similarity score that is measured by correlations in a media subspace between different poses on patch level. The media subspace is learned by Canonical Correlation Analysis (CCA) in order to maximize the intra-individual correlations.

A powerful tool to model non-rigid transformation, firstly proposed in computer graphics literature [27] and then applied to face recognition [26], is the 3DMM which has been introduced by Blanz and Vetter to model deformation of a face. The model is learnt analyzing the principal components

on a statistic of aligned 3D faces. In particular, in [26], Blanz and Vetter propose to recognize faces by fitting a 3DMM in the gallery and in the probe and using the retrieved coefficients as discriminative feature vector. In contrast to 3DMM, recently, authors in [70, 102] proposed an efficient way to estimate a 3D model from a single frontal image using their Generic Elastic Model (GEM). The GEM assumption is that the depth variation is not containing enough information if we consider a human face and the GEM model considers just a deformation on the XY plane as sufficient to obtain quite realistic 3D models. The method has been further improved considering diverse average values of depth per ethnic group [70]. Among all these papers, very recently, the attention has been moving on trying to normalize the pose of the profile face to a canonical frontal view. The paper in [7] is the first paper that reports the problem of landmark ambiguity. The approach synthesizes a frontal view from a profile one, exploiting a collection of set of manually labeled landmarks. They manually label one set of landmarks per pose to deal with self-occlusion. Face are normalized using a weak perspective pose estimation method and through a refining algorithm that extracts the boundary of the face. Boundary extraction is prone to fail when the background is not uniform, condition that typically occurs in video-surveillance imagery. Normalized faces are finally recognized using LGBP (Local Gabor Binary Pattern). Inspired by 3DMM, authors in [85] encode the pose variation of a test image in a linear combination of displacement fields, that they call Morphable Displacement Field. The approach is demonstrated to be robust and needs just eye-based alignment to process an image but the displacement field optimization is run for each image in the gallery, which is a strategy that does not scale for large galleries. Recently, in contrast to face normalization to a frontal view, authors in [134] tried also to modify a bank of Gabor Filters by localizing the filters in a precise manner optimizing a 3DMM on a probe image. In [111], Sharma *at al.* propose the Discriminant Multiple Coupled Latent Subspace framework. Similar to [84], the approach finds sets of projection directions for different poses such that the projected images of the same subject are maximally correlated in the latent space. Discriminant analysis with artificially simulated pose errors in the latent space makes it robust to small pose errors due to an incorrect pose estimate.

7.3 Automatically annotate and align 3D models

The dataset that was used to obtain a statistic of 3D faces is the Florence 2D/3D Face Dataset [11, 12] which collects both 2D imagery in the form of video sequences from a PTZ camera and 3D models recorded with a 3D scanner. We use the first raw frontal 3D model for each subject along with the associated texture. For more details about the database, please see [11].

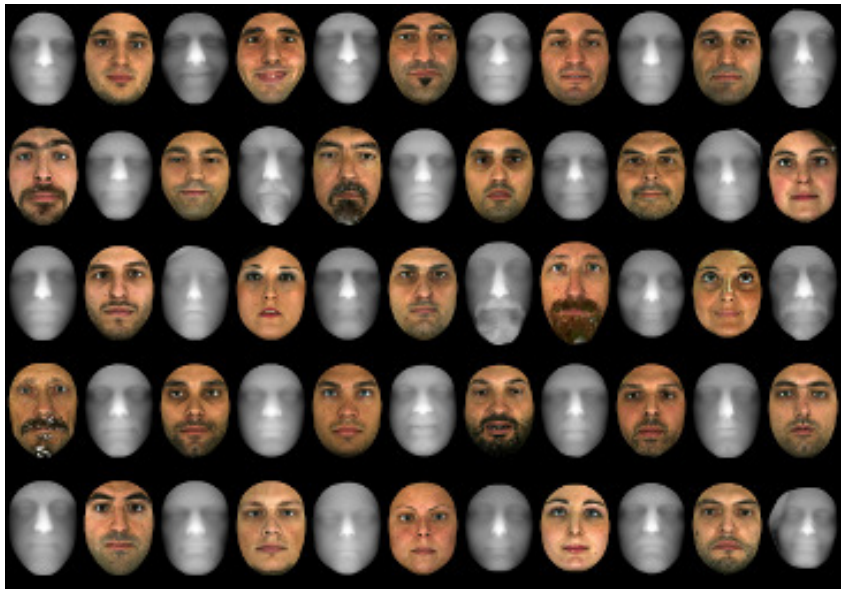


Figure 7.1: The face normalization process that brings all the meshes to a canonical frontal form: for each subject, the depth image and the rendered texture image are shown. Note that here the meshes do not have the per vertex alignment property.

Regarding the raw 3D models, we create a fully automated process to create canonical form of both the 3D models and the texture associated with the model. This permits us computing an average model that it is used to additionally parameterize all the models with the same number of vertices. This brings these benefits:

1. all the subjects are almost registered together,

2. the models are cleaned from noise considering hairs and ears and
3. finally the depth image and the texture image are perfectly registered together.

To achieve this we extract the coordinates of each 3D vertex along with its projection on the 2D image. In this chapter, we represent a 3D face as a matrix $\mathbf{V} \in \mathbb{R}^{3\times P}$ where each $\mathbf{V}_i = [x_i\ y_i\ z_i]^\top$ column represents a 3D vertex of the mesh and P is the number of points in the mesh. The texture of each model is represented as matrix $\mathbf{B} \in \mathbb{R}^{5\times P}$ where each $\mathbf{B}_i = [u_i\ v_i\ R_i\ G_i\ B_i]^\top$ column represents respectively the vertex projection on the texture $[u_i\ v_i]$ and the corresponding color value $[R_i\ G_i\ B_i]$.

In order to get the canonical form we use an iterative approach exploiting PCA, similar to [24, 94], and we process all the subject independently: at each iteration we compute the PCA of the covariance matrix of $\mathbf{V}$. We center the matrix by subtracting the mean $\mathbf{m} \in \mathbb{R}^3$ of the points (barycenter) and compute the covariance matrix as:

$$\mathbf{C} = \frac{1}{p}\sum_{i=1}^{p} \mathbf{V_i}\mathbf{V_i}^\top - \mathbf{m}\mathbf{m}^\top. \tag{7.1}$$

Performing PCA on this matrix, it gives us the rotation matrix $\mathbf{R}$ that align all the point cloud $\mathbf{V}$ on its principal axis. This method aligns all the faces in the same manner because the statistic of the face in the 3D space is approximatively the same for each subject.

The rotation matrix $\mathbf{R}$ corresponds to the eigenvector matrix extracted using PCA. At each iteration all the points in the matrix V are centered and rotated with the rotation matrix as:

$$\mathbf{V}' = \mathbf{R}(\mathbf{V} - \mathbf{m}) \tag{7.2}$$

Once the point cloud has been rotated, we proceed to coarsely localize the nose tip in order to remove the possible noise contained in the point cloud. The nose tip is found slicing the mesh on the YZ plane and detecting the nose tip as the maxiumum. Once the nose tip location is detected, we crop the 3D mesh by removing all the points that lie outside a ball centered in the nose with radius 100 mm. Then, in order to render a depth image and a 2D image, the algorithm proceeds to create a regular grid in a XY representation, that works as a support region, where both the z_i-value in $\mathbf{V}$ and the texture $[R_i\ G_i\ B_i]$ in $\mathbf{B}$ are interpolated with cubic interpolation

at a fixed sampling step (0.5 mm)[2]. The algorithm stops when a predefined number of iterations are reached or when the estimated rotation matrix is close to the identity matrix. The last condition usually holds. This overall process above gives the results shown in Fig. 7.1.

7.3.1 Building a generic 3D face model

Since now all the models are in the same canonical form and are aligned together is straightforward to build an average model that it will be used in the rest of the work as generic reference model. In order to proceed to average all the depth images and the texture, we repeat the algorithm of Section 7.3 with the same support region for all the subjects, in order to get the same image size. After that, we proceed to compute the mean on the support region, thus in XY representation, by averaging the z_i-value in $\mathbf{V}$ and the texture present in $\mathbf{B}$.

Once the average depth image is created, the generic face can be expressed in 3D representation.

7.3.2 Semantic Alignment and Landmarks Detection

All the models now are coarsely aligned but each vertex does not have the same semantic meaning along all the subjects. To overcome this, we parameterize each subject using the vertex of the generic model. We align each subject to the average face model using ICP with a 3D affine deformation that accounts for rotation, translation, size and shear as:

$$\min_{\mathbf{R},\mathbf{S},\mathbf{t},\mathbf{V}'} \sum_{i}^{P} \|\mathbf{RSV}_i + \mathbf{t} - \mathbf{m}_i\|_2 \tag{7.3}$$

being $\mathbf{R}$ an orthogonal matrix, $\mathbf{S}$ a diagonal scale matrix, $\mathbf{t}$ a translation vector. Considering that all the faces are already coarsely aligned, the ICP algorithm starts from a point near the optimum.

Given a subject, by using each point of the average model $\mathbf{m}$, we find the index of the closest point with euclidean distance between $\mathbf{m}$ and $\mathbf{V}'$, creating a collection of indexes that we use to index the aligned face model $\mathbf{V}'$ and its texture. In this way we transforme all the models to satisfy the

[2]This parameters can be tuned in order to get bigger images: the small is the sampling step, the bigger is the size of the image

per vertex alignment property, obtaining, considering just the shape, the matrix $\mathbf{S}$:

$$\mathbf{S} = \begin{bmatrix} \mathbf{V}_1^{(1)} & \mathbf{V}_1^{(2)} & \dots & \mathbf{V}_1^{(S)} \\ \mathbf{V}_2^{(1)} & \mathbf{V}_2^{(2)} & \dots & \mathbf{V}_2^{(S)} \\ \dots & \dots & \ddots & \dots \\ \mathbf{V}_P^{(1)} & \mathbf{V}_P^{(2)} & \dots & \mathbf{V}_P^{(S)} \end{bmatrix}. \tag{7.4}$$

The same holds for the texture part $\mathbf{T}$. Formally, from now on, we assume that $\mathbf{S}$ is centered on its empirical mean.

Similar to the dataset of Blanz and Vetter [26, 27], this brings two important benefits: (1) now each index of the mesh share the same semantic meaning across all the 3D faces (2) all the models have the same number of vertex.

Exploiting this property, it is possible to detect the reference 3D landmarks in a fully automatic way, by first detecting landmarks on the 2D rendered image for all the subjects; then computing an average of landmark detector responses and afterwards expressing the landmark locations in term of the 3D model, considering that the 2D image and the depth are perfectly aligned together.

Landmark Detection on 3D Models

The 3D face models share the per vertex alignment property. Exploiting this property, we design a way to automatically detect fiducial 3D landmarks on each model. To this end we rendered the texture in a frontal pose and we employ a landmark detector [142] that provides good recall in the face detection task and decent precision in the localization.

Once we extracted all the landmarks responses from all the frontal renderings, we average the responses taking the median. Our rendering software provides the ability of associating the 2D locations of a landmark with a 3D coordinates $\mathbf{V}_i = [x_i \ y_i \ z_i]^\top$ of the mean surface. We store the indexes $\mathbf{I}$ of these 3D labeled points that form the 3D reference landmarks. This procedure solves the issues reported in [7] of having landmarks manually annotated which can lead to inconsistent annotations. Even if we used just a set of frontal landmarks, we could potentially get 3D labeled landmarks in different poses, by just rendering the model in a given pose and repeat the above procedure.

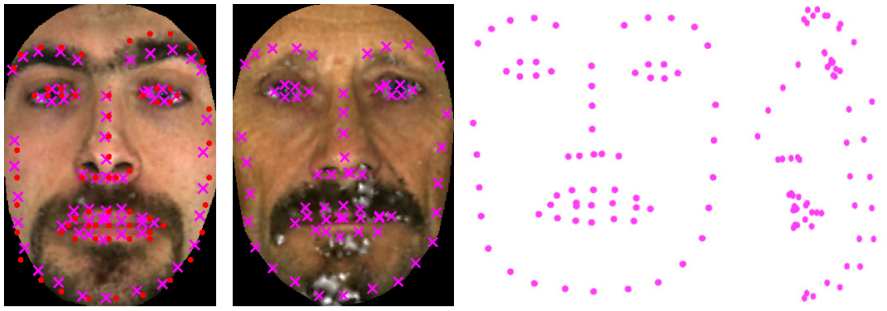


Figure 7.2: *Left:* In red the detected points on a rendering, while in magenta the correction that we get. The subject in the middle was not detected by the detector. *Right:* The 3D labeled landmarks.

Considering the per-vertex alignment property, we are able also to propagate the 3D landmarks to each subject in the dataset. Note that this is more robust then just detect on a face image and then back-project to each model, because by averaging the responses (1) we are accounting for some noisy responses going outside of the model in the subject having bear or mustaches (2) we are labeling subjects that have not been detected. See Fig. 7.2 for one example about the corrections explained above.

7.4 Learning 3D Deformation Components via 3DMM

The authors in [26] showed how to build a Gaussian model from a set of 3D aligned faces considering both the shape and the texture by using Principal Component Analysis (PCA) to obtain the principal components. We apply this approach on a dataset of 200 aligned 3D untextured faces taken from a commercial software that can produce virtual human faces (*virtualfaces*) and to our Florence Face database (*florencefaces*). Considering that the faces have a small number of vertices, we augment the mesh dimension performing a two-pass of loop-subdivision, that is able to maintain the alignment for each vertex. We stack all the linearized vertices in a matrix $\mathbf{S}$ where each row corresponds to a subject. We then proceed to label manually a set of indexes $\mathbf{I}$ on the average model $\mathbf{m} = \frac{\mathbf{1}}{\mathbf{S}} \sum_{\mathbf{i}}^{\mathbf{S}} \mathbf{S_i}$ that represents 3D reference landmarks and S is the number of models.

PCA factories $\mathbf{S}$ as follows:

$$\mathbf{S} = \mathbf{W}\ \mathbf{C}, \tag{7.5}$$

where $\mathbf{W} \in \mathbb{R}^{S \times K}$ represents the new dataset expressed in the new subspace and $\mathbf{C} \in \mathbb{R}^{K \times 3N}$ are the K components that correspond to the eigenvectors with $K \leq S - 1$. The benefit of using PCA is that each eigenvector has a eigenvalue which is a scalar, quantifying the amount of variance in that direction. In general the vector of eigenvalues is defined as $\boldsymbol{\sigma} = [\sigma_1 \dots \sigma_K]$ from the largest to the smallest one. This vector is important because gives a way to estimate of the probability density within a face space as $p(\boldsymbol{\alpha}) \sim \mathcal{N}\big(\mathbf{0}, \mathrm{diag}(\boldsymbol{\sigma}^2)\big)$.

Considering that N is very large, instead of estimating PCA using eigen-decomposition of the covariance matrix $\mathbf{S}\ \mathbf{S}^\top$, we rather use a more efficient way by performing Singular Value Decomposition (SVD) on $\mathbf{S}$ divided by $\sqrt{S-1}$. This procedure returns the matrix $\mathbf{C}$ and the corresponding eigenvalues as $\boldsymbol{\sigma}$.

In Fig. 7.3 are shown the deformation components of the shape and the texture when available of virtual faces and Florence face dataset. Once the deformation components are learnt, similarly to [26], we can easy synthesize a new face from:

$$\mathbf{S}' = \mathbf{m} + \sum_{k=1}^{K} \alpha_k\ \mathbf{C}_k. \tag{7.6}$$

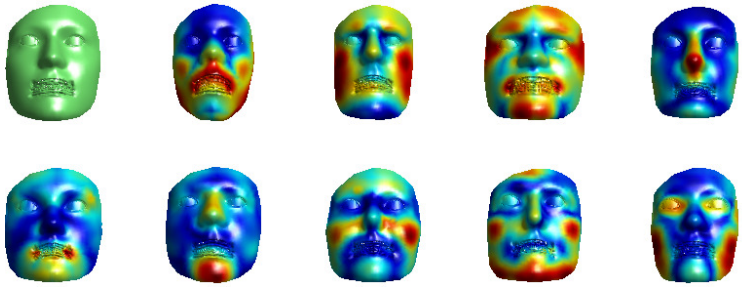
where

$$p(\boldsymbol{\alpha}) = p(\alpha_1, \dots, \alpha_K) \sim \exp\Big(\sum_{k=1}^{K} (\alpha_k / \sigma_k^2)\Big), \tag{7.7}$$

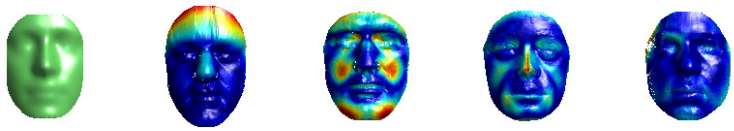
7.5 Fitting a Regularized 3DMM

In order to fit a 3DMM given a test image, we need firstly to get an estimate of the 3D pose in the image (rigid transformation) and then to fit the non-rigid transformation to the face (face model deformation). The cost function that we use it is simply the reprojection error of landmarks: as argued by [26], the use of this function could cause overfitting in the deformation process, leading to a surface which is not event close to a face[3].

[3]This happens considering that there are a lot of surfaces that can project on the image minimizing the reprojection error.



(a) Shape - virtualfaces



(b) Shape - florencefaces

(c) Texture - florencefaces

Figure 7.3: Deformation of the Components: *(first row:)* nine constructed components from left to right with virtual human faces. *(second row:)* four constructed components from Florence Faces. *(third row:)* four constructed components from Florence Faces of the texture. The average surface is always shown in the left-most box.

To overcome this, we propose to use a regularization similar to [134]: we formally address this problem as a Ridge Regression in which the coefficient energy of the 3DMM is proportionally limited by the prior given by the eigenvalues learned in Sec. 7.4. We proceed to explain firstly how the pose is estimated and then how the model is deformed.

7.5.1 3D Face Pose Estimation

In order to deal with face pose variation, it is necessary to establish correspondences between the labeled landmarks $\mathbf{m}(\mathbf{I})$ in the generic model and

the face framed in the image. To this end, we employ the landmark detector [142] that provides good recall in the face detection task and decent precision in the localization. We get an estimate of the pose using an affine camera model that maps each vertex of the model to the image. Under the affine camera model, the relation between the annotated points on the model $\mathbf{L} \doteq \mathbf{m}(\mathbf{I}) \in \mathbb{R}^{3\times N}$ and the detected points $\mathbf{l} \in \mathbb{R}^{2\times N}$ is the following:

$$\mathbf{l} = \mathbf{A}\ \mathbf{L} + \mathbf{t} \tag{7.8}$$

where $\mathbf{A}$ contains the affine camera parameters and $\mathbf{t} \in \mathbb{R}^{2\times N}$ is a translation on the image.

To recover these parameters, firstly we subtract the mean from each points set, then we recover the 2×3 affine matrix in a least square sense as $\mathbf{A} = \mathbf{l}\cdot\mathbf{L}^{\dagger}$, where $\mathbf{L}^{\dagger}$ is the pseudo-inverse matrix of $\mathbf{L}$. Secondly, we estimate the translation as $t = \mathbf{l} - \mathbf{A}\ \mathbf{L}$.

Furthermore, the matrix $\mathbf{A}$ can be decomposed with QR decomposition in two matrices: a matrix $\mathbf{S} \in \mathbb{R}^{2\times 3}$ that expresses the scale parameters along with the shear and another one, $\mathbf{R} \in \mathbb{R}^{3\times 3}$, that contains the 3D rotation parameters of the model with respect to the image. The final affine camera models is thus defined as:

$$\mathbf{l} = \mathbf{S}\ \mathbf{R}\ \mathbf{L} + \mathbf{t}. \tag{7.9}$$

Considering Eq. (7.8), it is possible to get an estimate of the pose considering the rotation matrix and to map each vertex of the generic model on the image.

7.5.2 Landmark Ambiguity Resolution

One of the problem reported in literature is the ambiguity in the landmark locations detected on an image w.r.t. the 3D reference landmarks in the model. The ambiguity does not hold if we consider a frontal face, but when the face undergoes large pose variations, the locations of detected landmarks become unstable especially if we consider fiducial landmarks on the boundaries like the jaws. This issue was reported in [7] and [134]: while the authors of [7] use a look-up table to access different 3D reference landmarks for each pose, the authors of [134] do not use at all the boundary landmarks.

In our approach we overcome this issue in this way: given a test image, we automatically select stable 3D landmarks from the ones present in the labeled

Algorithm 2: Landmark Ambiguity Resolution

Input: 3D landmarks $\mathbf{L}$, detected landmarks $\mathbf{l}$
Output: $\mathbf{S}$, $\mathbf{R}$, $\mathbf{t}$, $\mathbf{I_v} \subseteq \mathbf{I}$

1 Estimate pose $\mathbf{A}$ with Eq. (7.8) using all the indices $\mathbf{I}$.
2 Decompose $\mathbf{A}$ in order to get rotation matrix $\mathbf{R}$.
3 Estimate visibile landmarks as $\mathbf{I_v} = \text{HPR}(\mathbf{m}(\mathbf{I}), \mathbf{R})$.
4 Re-Estimate the pose $\mathbf{A}$ with Eq. (7.8) using $\mathbf{I_v}$.
5 Decompose $\mathbf{A}$ and return $\mathbf{S}$, $\mathbf{R}$, $\mathbf{t}$, $\mathbf{I_v} \subseteq \mathbf{I}$.

set $\mathbf{I}$. As the face undergoes self-occlusion caused by the pose, some 3D reference landmarks will not be visible. Our approach automatically selects the visible landmarks to use as reference using an Hidden Point Removal Operator (HPR) [79], and refines the pose using just a subset of indices $\mathbf{I_v} \subseteq \mathbf{I}$. The algorithm is reported in Alg. 2. This process gives a way to select stable landmarks and handles continuous pose variation and this information of visibilities will be used for recognition to select from the gallery the visible part of the face. We use all the landmarks for frontal gallery faces, omitting the boundary ones at recognition time.

7.5.3 Optimization

Given a test image we want to find the 3D face pose that give us a rigid transformation $\mathbf{P}$ and a non-rigid transformation in term of deformation components $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_K]$. Formally we want to optimize the following:

$$\min_{\mathbf{P},\boldsymbol{\alpha}} \left\| \mathbf{l} - \mathbf{P}\Big(\mathbf{m}(\mathbf{I_v}) + \sum_{k=1}^{K} \alpha_k \mathbf{C_k}(\mathbf{I_v})\Big) \right\|_2 + \lambda \, \|\boldsymbol{\sigma}^{-1} \cdot \boldsymbol{\alpha}\|_2 \tag{7.10}$$

ñ where $\boldsymbol{\sigma}$ is defined as in Sect. 7.4, $\mathbf{I_v}$ expresses the indices of the visible landmarks, $(\cdot)$ means the element-wise multiplication and λ is a scaling regularization parameter and has been set to 25. This values is set accordingly to the magnitude of eigenvectors.

We solve this problem by alternating between pose estimation and model coefficient estimation. We proceed to estimate the pose as report in Sect. 7.5.1 and then given the estimated pose $\mathbf{P}$, we solve for the coefficients and the

problem becomes:

$$\min_{\boldsymbol{\alpha}} \left\| \mathbf{l} - \mathbf{P}\mathbf{m}(\mathbf{I_v}) - \sum_{k=1}^{K} \mathbf{P}\alpha_k \mathbf{C_k}(\mathbf{I_v}) \right\|_2 + \lambda \, \|\boldsymbol{\sigma}^{-1} \cdot \boldsymbol{\alpha}\|_2. \tag{7.11}$$

By defining $\mathbf{X} \doteq \mathbf{l} - \mathbf{P}\ \mathbf{m}(\mathbf{I_v})$, $\mathbf{Y} \doteq \mathbf{P}\ \mathbf{C}(\mathbf{I_v})$ we get:

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \Big\| \underbrace{\mathbf{l} - \mathbf{P}\ \mathbf{m}(\mathbf{I_v})}_{X} - \sum_{k=1}^{K} \alpha_k \underbrace{\mathbf{P}\ \mathbf{C_k}(\mathbf{I_v})}_{Y} \Big\|_2 + \lambda \, \|\boldsymbol{\sigma}^{-1} \cdot \boldsymbol{\alpha}\|_2 = \\ = \min_{\boldsymbol{\alpha}} \left\| \mathbf{X} - \sum_{k=1}^{K} \alpha_k \mathbf{Y} \right\|_2 + \lambda \, \|\boldsymbol{\sigma}^{-1} \cdot \boldsymbol{\alpha}\|_2, \end{aligned} \tag{7.12}$$

and thus we can rewrite the system by linearizing the landmarks without changing the meaning, casting the problem as a regularized least squares one:

$$\min_{\boldsymbol{\alpha}} \|\mathbf{X} - \boldsymbol{\alpha}\mathbf{Y}\|_2 + \lambda \, \|\boldsymbol{\sigma}^{-\mathbf{1}} \cdot \boldsymbol{\alpha}\|_2 \tag{7.13}$$

which is analytically solved as a Ridge Regression where each component is weighted using the inverse of σ_k taken from the $\boldsymbol{\sigma}$. The non-rigid coefficients are retrieved as:

$$\boldsymbol{\alpha} = (\mathbf{Y}^T\mathbf{Y} + \lambda \, \mathrm{diag}(\boldsymbol{\sigma}^{-1}))^{-1}\mathbf{Y}^T\mathbf{X}. \tag{7.14}$$

It is important to notice that the number of unknowns in the system are the number of components K while the constraints are the number of landmarks $2N$. So it is important that $K \leq 2N$ to maintain the problem not underdetermined. Fig. 7.4 shows the geometrical meaning of the optimization process along with the recovered shapes with and without regularization. The figure shows just a landmark for the sake of clarity: the projected components in blue indicates the directions on which deform the model in order to minimize the reprojection error.

Once the new shape is obtained using the new $\boldsymbol{\alpha}$ as detailed in Eq. (7.6) , we proceed to perform a finale estimation of the pose $\mathbf{P'}$ with the new shape $\mathbf{S'}$ and additionally to select the visible indices of vertices of the entire shape $\mathbf{J}_v$. In the following Section, we use these two last estimates to synthesize a frontal view from a profile face accounting for rigid and non-rigid transformation.

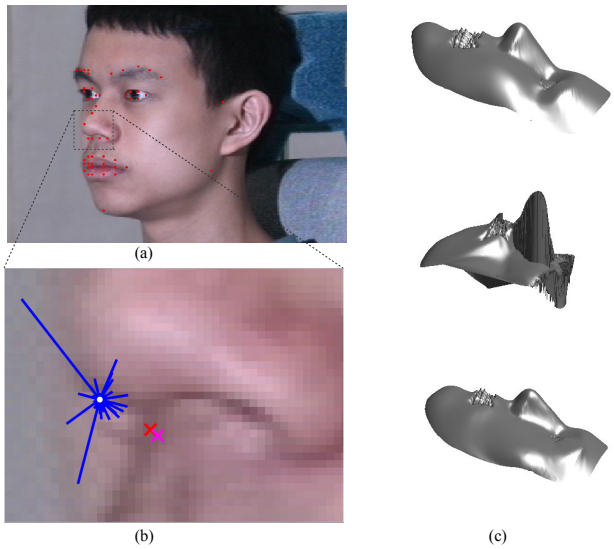


Figure 7.4: (a) Detected landmarks (b) Geometric meaning of deformation components: red cross is the detected landmark while the white dot is the projected landmark from the generic model (the initial estimation). The deformation components in blue indicate the directions on which deform the model, while the coefficients are the relative magnitude. Once the model is fit, the landmark projects on the magenta cross. (c) Respectively from top to bottom: average model, estimate of shape without regularization and finally with the proposed regularization. Note how the chin and the nose have changed.

7.6 Recognition by Localizing Local Binary Patterns on the Deformed Vertices

In this Section we describe our face recognition method across a range of facial poses. Considering the result achieved in the previous Sections, differently from recent approaches likes [7,8,134], our method can handle continuos pose variations and does not require any manual labeling of data except for the 3D landmarks in the average model.

Our approach supposes to process the gallery by fitting the 3DMM for each subject and by extracting LBP histogram on a window localized on

each vertex of the deformed model. Furthermore in order to reduce the dimensionality of the descriptor, we sampled equally the vertices. From our experimental results we deduced that a 2D Gaussian low-pass filter applied on the gallery images enhances the performance. Thus we apply it with a size of 5×5 and a standard deviation of 0.9 on the frontal face rendering. Once the LBP histograms are extracted, they are then stacked together to form a unique descriptor similar to [4].

Defining a LBP on the vertex has several benefits w.r.t. the state-of-the-art method [4] that divides the image with a regular uniform grid and extract LBP histogram in each cell of the grid. These benefits are the following:

- this makes the feature vector independent of the image size.
- considering that the deformed model has been optimized to fit the face, the LBP are better localized.
- it gives a straightforward way to restrict the feature vector to those parts which are not visible considering a self-occluding face.

Our method to recognize a face is the following. Given a test image, once we have an estimate of pose $\mathbf{P}'$ and a regularized shape $\mathbf{S}'$, we proceed to render a frontal view and we sample LBP on just a uniform subset of the deformed visible vertices. So in this case when the face undergoes self-occlusion, the feature dimension of the query will be less than the one in the gallery. To this end, we exploit the visible indices $\mathbf{J}_v$ and select the part of the feature vector in the gallery that corresponds to these indices. Once we a have a face descriptor $\mathbf{f} \in \mathbb{R}^{F(\mathbf{J}_v)}$ for a query image, we simply apply a Nearest Neighbor (NN) algorithm to select the closest feature from the ones in the gallery $\mathbf{G} \in \mathbb{R}^{F(\mathbf{J}_v) \times N_s}$ where N_s is the number of subject in the gallery and $F(\mathbf{J}_v)$ is the feature dimension that arises considering each time the visible landmarks. Thus our recognition rule simply is:

$$id(\mathbf{f}) = \arg\min_i \|\mathbf{f} - \mathbf{G}_i\|_2 \tag{7.15}$$

Each frontal face image is rendered on the XY plane considering the deformed shape, interpolating the RGB values sampled from the non-frontal face on an uniform grid with natural neighbor interpolation. This generally gives a face image size of about 120×160 pixels. On each vertex we sample a patch of size 11 pixels and each LBP histogram is quantized in 58 bins, where two bins account for non-uniform binary patterns and the remaining count the uniform binary patterns.

7.7 Experimental Results

(a) (b) (c) (d) (e) (f)

Figure 7.5: Qualitative examples: *(a)* the probe and the gallery image cropped with similarity alignment. *(b-f)* From left to right we show frontal renderings with the following: 2D similarity method; generic model using all the landmarks; generic model with only selected landmarks; morphed model without regularization; morphed model with regularization.

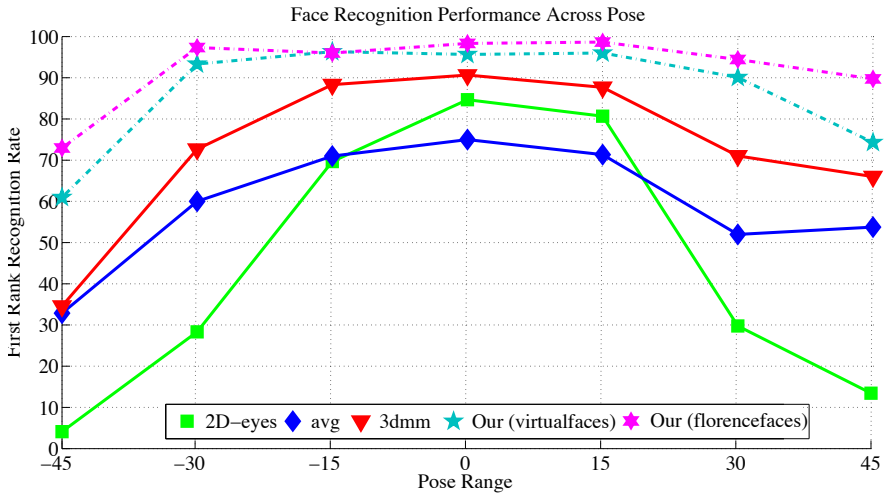
In this section we report the experiment results obtained using our method, comparing the performance figures with the approach of [4] considering various alignment modalities such as:

- *2d-eyes*: 2D similarity alignment approach based on eyes-mouth locations. The similarity is defined by the triangle formed by the eyes and the mouth that maps into a triangle in a template image of size 200×240;
- *avg*: render a frontal image by pose normalization using a 3D average model and a final 2D similarity alignment.
- *3dmm*: similar to the previous one: obtaining a frontal image by pose normalization using a 3DMM and performing a final 2D similarity alignment.

All the approaches use the same landmark detector, that is the one used specified in Sect. 7.5.1. If we apply the approach of [4] on the frontal renderings, these have different size and the above approach does not provide a way to extract it with a fixed length. To overcome this, we project on the rendered face the 3D landmarks and by selecting the triangle connecting the eyes and the mouth, we perform a final 2D similarity alignment that gives an image of the same size, irrespective of the rendering size. Considering these baselines, for the face poses at $\pm 45°$we use just the left/right part of the face for matching.

In addition to these baselines, we compare with state-of-the-art results in [7, 8, 85, 134, 136] on a regular dataset used in literature: we test our approach on CMU Multi-PIE, expressing the performance by principally

reporting the recognition rate at first rank under a range of facial poses. However in order to show the potentiality of our method we show also the full CMC (Cumulative Matching Characteristic) curves along with the nAUC (normalized Area Under the Curve).



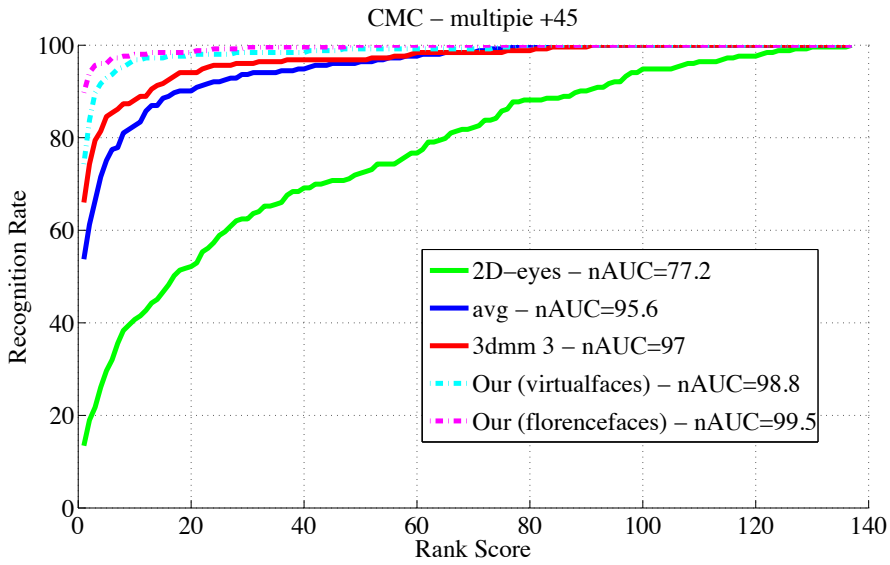
(a)

Figure 7.6: The first rank recognition rate as a function of the facial pose on the CMU-MPIE dataset.

CMU Multi-PIE is the most recent of controlled face databases and our experiments on this will facilitate the comparison with future methods. Moreover this dataset is very complete because the subject are framed under every possible conditions: by varying the pose, illumination and the facial expression.

We recreate exactly the experiment settings of the approach [7] considering 137 subjects (subject from 201 to 346) with neutral expression from all 4 sessions at 7 different poses, with illumination that is frontal with respect to the face (see labels in Tab. 7.1(b)).

We use the frontal image (label pose 051) from the earliest session for each subject as the gallery image (137 total) and all of the remaining per subject as the probe set. The dimension of the probe set is $1,963$ images. Note that differently from other approaches that trained PCA and LDA [85, 134] or learned the landmark detector on the first 200 subjects [7], we



(a)

Figure 7.7: CMC curves with probe faces showing a pose of 45°along with the nAUC.

did not use these data nor use PCA/LDA. In Fig. 7.5 we show some face rectification for the subject 201 when the face undergoes a pose variation in yaw of $-45°$: on the left we report the probe profile image along with the gallery image aligned with 2D similarity. Then from left to right the rendered frontal image, respectively using the 2D similarity method (*2D-eyes*); the generic model using all the landmarks (*avg*); the generic model with only selected landmarks; the morphed models without regularization (*3Dmm*); and finally the morphed models with regularization. In Fig. 7.6 we show the face recognition performance across pose at first rank comparing the baselines. Considering LBP on a regular grid [4], from our experiment evaluation, it arises that even if the face is normalized to a frontal pose, they still do not give compelling performances. They do provide better performance then 2D similarity but our approach using LBP localized with deformation components outperforms both. Moreover we experimented that the proposed approach is more discriminative if it uses 3D real face models *(florencefaces)* than virtual human faces *(virtualfaces)*: to this end we report

Pose	-45°	-30°	-15°	0°	+15°	+30°	+45°	Mean
Label	*080_05*	*130_06*	*140_06*	*051_07*	*050_08*	*041_08*	*190_08*	–
LGBP [136]	37.7	62.5	77.0	92.6	83.0	59.2	36.1	64.0
BMVC05 [8]	43.8	83.3	94.0	96.3	94.7	70.0	41.2	74.8
ICCV11 [7]	74.1	91.0	95.7	96.9	95.7	89.5	74.8	87.7
ECCV12 s1 [85]	78.7	94.0	99.0	–	98.7	92.2	81.8	90.7
ECCV12 s2 [85]	84.7	95.0	**99.3**	–	99.0	92.9	85.2	92.7
CVIU12 [111]	**84.8**	96.6	99.2	–	**99.2**	**96.2**	89.0	**94.1**
Ours *(virtualfaces)*	61.0	93.3	96.3	95.6	96.0	90.0	74.3	86.6
Ours *(florencefaces)*	72.9	**97.3**	96.0	**98.3**	98.7	94.4	**89.7**	92.5

Table 7.1: Pose-wise first-rank recognition rates (%). Bold means best performance.

also the performance figure of our approach when it uses the 3D models of the Florence Faces dataset [11]. This can be observed also in Tab. 7.1 in which we present also a comparison with the state-of-the-art: it is shown a pose-wise breakdown of recognition rate at first rank against recent methods. From this comparison our approach shows comparable results against the state-of-the-art and in some case it reports better performance f.e. for poses at $\{-30, 0, +45\}^\circ$.

Chapter 8

Conclusions and Outlook

8.1 Concluding Remarks

In this thesis we have presented various methods that interpret video content in order to automatically analyze people in term of trajectories and associate people across camera using both body appearance and face biometric information. The analysis is carried out automatically by interpreting a video stream and is assisted by 3D data. The analysis is split in two different semantic layers: a high level part, in which the thesis proposes to observe people by extracting their paths; and a low level layer, which is focused on faces, the most promising and non-intrusive biometric. Depending on these two parts, different aspects of humans are captured.

8.2 Summary of Contributions

This thesis makes several contributions to the field of video understanding and computer vision. Our effort has been dedicated to finding new algorithms and practical methods that in a near future will be implemented and will lead to the state-of-the art in video surveillance applications and video analytics software. Though we did not address in this thesis the technical and practical problems concerned with the implementation of this kind of systems, from a theoretical point of view, this thesis has opened new futures in the field of the video analytics. Regarding the contributions of this work, we can summarize these in the following:

- A new approach for multi-person tracking in wide areas with a robotic PTZ sensor has been proposed in contrast to the current applications that rely on fixed camera and background subtraction algorithms (Chapter 2). The method is designed to be efficient while it has been demonstrated that it can achieve promising performance compared to the current state-of-the-art. The method is able to follow multiple persons and, even if the sensor is moved, it is able to extract long trajectories in world coordinates. This is possible with a supervised procedure that must be performed just once at setup time.
- The approach proposed in Chapter 2 cannot deal with a person that turns at a corner and cannot handoff between two non-overlapping cameras. To this end, the previous method has been extended by matching person bodies across cameras, in order to link the previous trajectories across a wide area (Chapter 3).
- We argue that a significant contribution to the analysis of people is given by biometric details. Thus another contribution of this thesis is to point out how to use the face biometric in order to recognize people from still images and videos. The answer is to improve face recognition algorithms to collect relevant face imagery (Chapters 4 and 5) and to simplify the enrollment process, thus using a single 3D textured model per person (Chapter 6), or to handle large pose variations as typically occurs in video surveillance scenarios (Chapter 7).
- A contribution that passes through all the Chapters is to show how 3D data can assist this automatic analysis of video either on motion analysis part either on face recognition algorithms.

8.3 Impact of the Proposed Research

We believe that the proposed research will have a high impact in the fields of video surveillance applications and video analytics. For example, our work on "Motion Analysis" will lay the foundations for next-generation surveillance systems that must monitor people in critical zones; it could be useful in the shopping centers if entrepreneurs want to have coherent statistics about the behavior of their clients. Moreover these methods could also be exploited in the sports domain (i.e. football player tracking) to improve statistical analysis of players or to improve the entertainment for the final user.

Regarding our work on "Face Analysis", we could image practical appli-

cation of impact if someone needs to search for an identity over a large face database having a face captured in critical conditions: a practical example is the beneficial impact that this research will have on the toolkits that forensic investigators use after some disasters like what happened in London in 2005 or at the Boston Marathon in 2013[1].

[1]For a real example please see the call for tips at `http://www.fbi.gov/news/updates-on-investigation-into-multiple-explosions-in-boston/photos`, where you can find also some face imagery taken from a real scenario.

Appendix A

Other Datasets

A.1 MICC PTZ Dataset

Sequence	*Seq.#1*	*Seq.#2*	*Seq.#3*	*Seq.#4*
Name	"Long"	"Focus"	"Dense"	"Rapid Motion"
#Frames	782	530	1750	600
Resolution	320×240	368×276	320×240	320×240
#Keyframes	269	144	190	140
#Zoom Levels	4	3	4	1
Density	Low	Medium	High	High
Place/Size	Out/Wide	Out/Wide	Out/Wide	Indoor/Small
Scene Texture	Weak	Weak	Weak	Good
Illumination	Natural	Natural	Natural	Artificial

Table A.1: Overview of the MICC PTZ dataset.

Standard public datasets do not contain sequences with a moving and zooming sensor (PTZ) and the relative calibration data, except for the UBC Hockey dataset. For this reason to test the performance of the proposed method we have collected and published a new PTZ dataset with different viewpoints and scenarios. We remark that all the sequences were acquired and processed with *one moving* camera. In particular, the outdoor sequences are acquired with a "PTZ Sony SNC-RZ30" while the indoor sequence is acquired with a "PTZ Axis Q6032-E". The map is learned by sampling the observed scene with pan, tilt angles taken respectively at about $20, 10$ degrees

between each keyframes for all the sequences. Both cameras allow to inquire the current motor odometry independently from the captured image.

The proposed dataset is composed by four sequences:

- Seq.#1 "Long": outdoor, low density of targets and long trajectory. Targets moves for about 60 meters in a parking area.
- Seq.#2 "Focus": outdoor, medium density of targets. The camera performs high zoom-in to focus on one of the tracked target during the sequence and returns on all targets after few frames.
- Seq.#3 "Dense": outdoor, high density of targets. Many targets move randomly in the area and leave and enter continuously the field of view.
- Seq.#4 "Rapid Motion": indoor, high density of targets with abrupt camera motion. Many targets moves contemporary in the scene while the camera is performing a rapid patrolling and zoom.

These PTZ sequences provide a total of 9,685 annotations for 3,662 frames.

(a) Seq.#1 Long

(b) Seq.#2 Focus

(c) Seq.#3 Dense

(d) Seq.#4 Rapid Motion

Figure A.1: Sample frames from the MICC PTZ Dataset. Each row represents a sequence.

A.2 Face Logging Dataset

Face logging dataset consists of a varying number of sequences taken at a frame rate of about 20 fps at different resolutions (see figure A.2).

The sequences in Dataset #1 were recorded so as to cover a wide variety of circumstances and to give a qualitative impression of the performance of our face logging system. For example, some of them are taken from a typical door or corridor surveillance vantage point. In these cases, it is quite easy to acquire good images of target faces. While other sequences are taken in a wide-angle indoor environment where faces appear at a distance and thus the resolution is very poor.

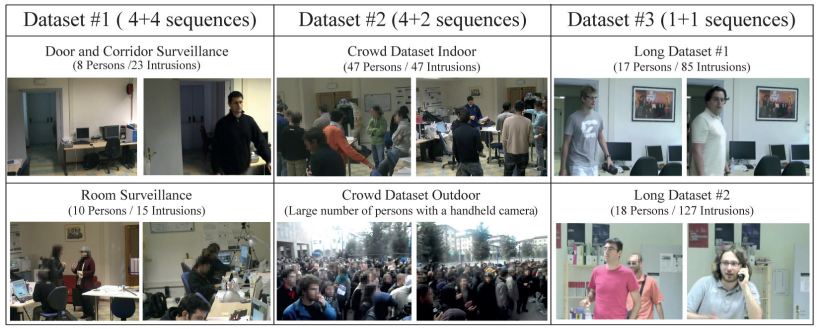


Figure A.2: Some frames from each sequence of the dataset.

Datasets #2 and #3 are specifically designed to provide a quantitative evaluation of face logging systems and to probe their effectiveness and robustness over an entire workday. In Dataset #2 several sequences were taken in very crowded environments such as would occur in real life video surveillance scenarios, both indoor and outdoor, where occlusions, cast shadows, and frequent changes of the head pose are very common. This dataset was collected specifically to evaluate the effectiveness of face logging systems under realistic, video surveillance scenarios. Sequences are taken with a varying number of people (from about ten to more than one hundred persons) and are composed of about 86 minutes of video of critical conditions in which there may be frequent identity mismatches due to the proximity of people.

Dataset #3 is composed of two sequences, each of about five hours, taken in an indoor environment. This dataset was collected to evaluate the long-term robustness of face logging systems.

Appendix B

Publications and Projects

Publications

This research activity has led to several publications in international journals and conferences. These are summarized below. The underlined name indicates the main investigator. If more than one name is underlined, that means equal contribution to the paper.

Submitted

- G. Lisanti, I. Masi, A. D. Bagdanov and A. Del Bimbo, "Person Re-identification by Iterative Re-weighted Sparse Ranking" in *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **(Under Major Revision)**.
- I. Masi, C. Ferrari, A. Del Bimbo, G. Medioni, "Pose Independent Face Recognition by Localizing Local Binary Patterns via Deformation Components" in *International Conference on Pattern Recognition 2014*, **(Under Review)**.
- P. Salvagnini, F. Pernici, M. Cristani, G. Lisanti, I. Masi, A. Del Bimbo and V. Murino, "Information Theoretic Sensor Management for Multi-Target Tracking with a Single Pan-Tilt-Zoom Camera" in *IEEE Winter Applications of Computer Vision Conference 2014*, **(To Appear)**.

International Journals

- A. D. Bagdanov, A. Del Bimbo, F. Dini, G. Lisanti and I. Masi, "Posterity Logging of Face Imagery for Video Surveillance" in *IEEE Multimedia 2012*.

International Conferences, Workshops and Demos

- Andrew D. Bagdanov, Alberto Del Bimbo, Dario Di Fina, Svebor Karaman, Giuseppe Lisanti, Iacopo Masi, "Multi-Target Data Association using Sparse Reconstruction" in *Proc. of International Conference on Image Analysis and Processing (ICIAP)*, Naples, Italy, 2013.
- I. Masi, G. Lisanti, A. D. Bagdanov, P. Pala and A. Del Bimbo, "Using 3D Models to Recognize 2D Faces in the Wild" in *Proc. of CVPR Int'l Workshop on Socially Intelligent Surveillance and Monitoring*, Portland, USA, 2013.
- A. D. Bagdanov, A. Del Bimbo and I. Masi, "Florence faces: a dataset supporting 2D/3D face recognition", in *IEEE 5th International Symposium on Communications, Control and Signal Processing*, Roma, Italy, 2012.
- A. D. Bagdanov, A. Del Bimbo, G. Lisanti and I. Masi, "Multi-pose Face Detection for Accurate Face Logging", in *International Conference on Pattern Recognition*, Tsukuba Science City, Japan, 2012.
- A. D. Bagdanov, A. Del Bimbo and I. Masi, "The Florence 2D/3D Hybrid Face Dataset", in *Joint ACM Workshop on Human Gesture and Behavior Understanding (J-HGBU'11) ACM Multimedia Workshop 2011*, Arizona, USA, 2011.
- A. Del Bimbo, G. Lisanti, I. Masi, F. Pernici. "Continuous Recovery for Real Time Pan Tilt Zoom Localization and Mapping", in *Proc. of 2011 IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS 2011)*, Klagenfurt, Austria, 2011.
- A. Del Bimbo, F. Dini, G. Lisanti, I. Masi and F. Pernici, "3D Multiple Target Tracking and Face Pose Estimation with a Rotating and Zooming Camera", in *ECCV – European Conference on Computer Vision, Demo Session*, Heraklion, Crete and Greece, 2010.
- A. Del Bimbo, G. Lisanti, I. Masi, F. Pernici. "Person Detection using Temporal and Geometric Context with a Pan Tilt Zoom Camera", in *Proc. of International Conference on Pattern Recognition (ICPR)*, Istanbul, Turkey, 2010.
- A. Del Bimbo, G. Lisanti, I. Masi, F. Pernici. "Device-Tagged Feature-based Localization and Mapping of Wide Areas with a PTZ Camera", in *Proc. of CVPR International Workshop on Socially Intelligent Surveillance and Monitoring (SISM)*, San Francisco, USA, 2010.

The author's bibliometric indices are the following: H-index = 3; the total number of citations = 15 (source: Google Scholar on December 9, 2013).

Projects

During the Ph.D., I have personally addressed some transfer projects with companies or helped in other projects. The projects that I was involved with are the following:

- PAR-FAS Local Project, Intermodal System Integrated for Security and Signaling on Rail.
- Image Stitcher software in C++ and ad-hoc API in C# for Cultural Heritage for the local company Culturanuova s.r.l.
- "Video analysis software for security applications" is an ongoing technology transfer and research project between Thales Italy and the Media Integration and Communication Center. The result of this research is a prototype that was presented as Very Important Demo at the Thales Technoday in Paris, France in 2011.
- Transfer project for Zucchetti Centro Sistemi S.p.A.: we developed a method that localizes a wheeled mobile robot observed from multiple ceiling cameras in real-time and drives the robot over a path in a large environment with a pure pursuit controller, achieving less then 5 pixels on cross track error.

Bibliography

[1] "http://www.micc.unifi.it/vim/datasets/mtt/." 31

[2] 3dMD the world leader in ultra-fast 3D surfacing imaging for healthcare, "3dMD face system," http://www.3dmd.com/. [Online]. Available: http://www.3dmd.com/ 100

[3] A. F. Abate, M. Nappi, D. Riccio, and G. Sabatino, "2D and 3D face recognition: A survey," *Pattern Recogn. Lett.*, vol. 28, no. 14, pp. 1885–1906, Oct. 2007. 108

[4] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 12, pp. 2037 –2041, dec. 2006. 10, 120, 134, 135, 137

[5] J. G. Allen, R. Y. D. Xu, and J. S. Jin, "Object tracking using camshift algorithm and multiple quantized feature spaces," in *Proceedings of VIP '05*. Australian Computer Society, Inc., 2005. 80, 81

[6] A. Ashraf, S. Lucey, and T. Chen, "Learning patch correspondences for improved viewpoint invariant face recognition," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008, pp. 1–8. 121

[7] A. Asthana, T. Marks, M. Jones, K. Tieu, and M. Rohith, "Fully automatic pose-invariant face recognition via 3d pose normalization," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, nov. 2011, pp. 937 –944. 10, 120, 122, 126, 130, 133, 135, 136, 138

[8] A. Asthana, T. K. Marks, M. J. Jones, K. H. Tieu, and M. Rohith, "Fully automatic pose-invariant face recognition via 3d pose normalization," in *Proceedings of the 2011 International Conference on Computer Vision*, ser. ICCV '11, 2011, pp. 937–944. 133, 135, 138

[9] T. Avraham, I. Gurvich, M. Lindenbaum, and S. Markovitch, "Learning implicit transfer for person re-identification," in *Proceedings of ECCV - Workshops and Demonstrations*, 2012, pp. 381–390. 48

[10] A. D. Bagdanov, A. Del Bimbo, F. Dini, G. Lisanti, and I. Masi, "Posterity logging of face imagery for video surveillance," *IEEE Multimedia*, vol. 19, no. 4, pp. 48–59, oct-dec 2012. 110, 116

[11] A. D. Bagdanov, A. Del Bimbo, and I. Masi, "Florence faces: a dataset supporting 2d/3d face recognition," in *Proc. of Int. Symposium on Communication Control and Signal Processing (ISCCSP)*, Rome, Italy, 2012. 113, 123, 138

[12] A. D. Bagdanov, I. Masi, and A. Del Bimbo, "The florence 2d/3d hybrid face datset," in *Proc. of ACM Multimedia Int.'l Workshop on Multimedia access to 3D Human Objects (MA3HO'11)*, ACM. ACM Press, December 2011. [Online]. Available: http://www.micc.unifi.it/publications/2011/BMD11 123

[13] S. Bak, G. Charpiat, E. Corvee, F. Bremond, and M. Thonnat, "Learning to Match Appearances by Correlations in a Covariance Metric Space," in *European Conference on Computer Vision*. Springer, 2012. 48, 60, 65

[14] S. Bak, E. Corvee, F. Bremond, and T. Monique, "Multiple-shot Human Re-Identification by Mean Riemannian Covariance Grid," in *2011 8th IEEE International Conference on Advanced Video and Signal-Based Surveillancei (AVSS)*, Aug. 2011, p. 6. 47, 60, 65, 67

[15] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Person Re-identification Using Spatial Covariance Regions of Human Body Parts," in *7th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AVSS - 2010*, Aug. 2010. 47, 60, 65

[16] L. Barceló, X. Binefa, and J. R. Kender, "Robust methods and representations for soccer player tracking and collision resolution," in *Proceedings of the International Conference on Image and Video Retrieval*, 2005. 18

[17] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded Up Robust Features." in *Proceedings of the European conference on Computer vision*, 2006. 20, 22

[18] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino, "Multiple-shot person re-identification by hpe signature," in *20th International Conference on Pattern Recognition (ICPR)*, 2010, pp. 1413 –1416. 44, 47, 60, 65, 67

[19] L. Bazzani, M. Cristani, A. Perina, and V. Murino, "Multiple-shot person re-identification by chromatic and epitomic analyses," *Pattern Recognition Letters*, 2011. 44, 47, 60, 65, 67

[20] ——, "Multiple-shot person re-identification by chromatic and epitomic analyses," *Pattern Recognition Letters*, vol. 33, no. 7, pp. 898 – 903, 2012. 116

[21] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 28

[22] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the CLEAR MOT metrics," *Journal on Image and Video Processing*, vol. 2008, Jan. 2008. 34

[23] S. Berretti, A. Del Bimbo, and P. Pala, "3D face recognition using isogeodesic stripes," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 32, no. 12, pp. 2162–2177, December 2010. [Online]. Available: http://www.micc.unifi.it/publications/2010/BDP10 98

[24] ——, "3d face recognition using isogeodesic stripes," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 32, no. 12, pp. 2162–2177, December 2010. 124

[25] S. T. Birchfield and S. Rangarajan, "Spatiograms versus histograms for region–based tracking," 2005. 29

[26] V. Blanz and T. Vetter, "Face recognition based on fitting a 3d morphable model," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 9, pp. 1063–1074, 2003. 121, 122, 126, 127, 128

[27] ——, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, ser. SIGGRAPH '99, 1999, pp. 187–194. 121, 126

[28] ——, "Face recognition based on fitting a 3d morphable model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, p. 2003, 2003. 108

[29] M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Online multi-person tracking-by-detection from a single, uncalibrated camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 99, p. 1, 2010. 31, 32, 39, 40

[30] M. D. Breitenstein, D. Kuettel, T. Weise, L. V. Gool, and H. Pfister, "Real-time face pose estimation from single range images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, June 2008. 98

[31] M. Brown and D. Lowe, "Recognising panoramas," in *Proceedings of the International Conference on Computer Vision*, 2003. 16, 20, 22

[32] E. Candès, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted l1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, pp. 877–905, 2008. 55

[33] K. I. Chang, K. W. Bowyer, and P. J. Flynn, "Multi-modal 2d and 3d biometrics for face recognition," in *Proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, 2003, pp. 187–. 108

[34] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *British Machine Vision Conference (BMVC)*, 2011. 47, 60, 65, 67

[35] Chinese Academy of Sciences' Institute of Automation (CASIA), "Casia-3d facev1," http://biometrics.idealtest.org/. [Online]. Available: http://biometrics.idealtest.org/ 96, 97, 101

[36] W. Choi, C. Pantofaru, and S. Savarese, "A general framework for tracking multiple people from a moving camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2012. 31, 32, 37, 38, 39, 40

[37] R. G. Cinbis, J. Verbeek, and C. Schmid, "Image categorization using Fisher kernels of non-iid image models," in *IEEE Conference on Computer Vision & Pattern Recognition, Jun 2012*, Providence, United States, Apr. 2012. 50

[38] J. Civera, A. J. Davison, J. A. Magallon, and J. M. M. Montiel, "Drift-free real-time sequential mosaicing," *International Journal of Computer Vision*, vol. 81, no. 2, pp. 128–137, February 2009. 16

[39] A. Criminisi, I. Reid, and A. Zisserman, "A plane measuring device," *Image and Vision Computing*, vol. 17, no. 8, pp. 625–634, 1999. 25

[40] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005. 14, 27, 28, 32, 33, 35, 50

[41] L. de Agapito, E. Hayman, and I. D. Reid., "Self-calibration of rotating and zooming cameras." *International Journal of Computer Vision*, vol. 45, no. 2, November 2001. 16

[42] A. Del Bimbo and F. Dini, "Particle filter-based visual tracking with a first order dynamic model and uncertainty adaptation," *Computer Vision and Image Understanding*, vol. 115, no. 6, pp. 771 – 786, 2011. 80, 81, 82

[43] A. Del Bimbo, F. Dini, and G. Lisanti, "A real time solution for face logging," in *Proceedings of the International Conference on Imaging for Crime Detection and Prevention*, London, United Kingdom, 2009. 78

[44] A. Del Bimbo, F. Dini, G. Lisanti, and F. Pernici, "Exploiting distinctive visual landmark maps in pan-tilt-zoom camera networks," *Computer Vision and Image Understanding (CVIU)*, vol. 114, no. 6, pp. 611 – 623, 2010. 94

[45] A. Del Bimbo, G. Lisanti, I. Masi, and F. Pernici, "Device-tagged feature-based localization and mapping of wide areas with a ptz camera," in *Proceedings of CVPR International Workshop on Socially Intelligent Surveillance and Monitoring*, 2010. 8, 19

[46] A. Del Bimbo, G. Lisanti, and F. Pernici, "Scale invariant 3D multi-person tracking using a base set of bundle adjusted visual landmarks," in *Proceedings of ICCV International Workshop on Visual Surveillance*, 2009. 19, 27

[47] V. Delaitre, D. F. Fouhey, I. Laptev, J. Sivic, A. Gupta, and A. A. Efros, "Scene semantics from long-term observation of people," in *Proceedings of the 12th European Conference on Computer Vision - Volume Part VI*, ser. ECCV'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 284–298. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-33783-3_21 1

[48] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," in *Proceedings of the 10th Asian conference on Computer vision - Volume Part IV*, ser. ACCV'10, 2011, pp. 501–512. 48

[49] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *CVPR*, June 2009. 14

[50] E. Elhamifar and R. Vidal, "Robust classification using structured sparse representation," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '11, 2011, pp. 1873–1879. 109

[51] G. Fanelli, J. Gall, and L. van Gool, "Real time head pose estimation with random regression forests," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*, June 2011. 98

[52] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010. 44, 47, 60, 63, 65, 67

[53] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *Int. J. Comput. Vision*, vol. 61, no. 1, pp. 55–79, Jan. 2005. 47

[54] P. Felzenszwalb, R. Girshick, and D. McAllester, "Cascade object detection with deformable part models," in *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 2010. 14, 32, 35, 37

[55] R. J. Fitzgerald, "Pack biases and coalescence with probabilistic data association." *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-21, 1985. 30

[56] P.-E. Forssen, "Maximally stable colour regions for recognition and matching," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, june 2007, pp. 1 –8. 47

[57] A. Fourney and R. Laganiere, "Constructing face image logs that are both complete and concise," in *Proceedings of the Canadian Conference on Computer and Robot Vision*, 2007. 76, 83, 85, 86

[58] H. Fronthaler, K. Kollreider, and J. Bigun, "Automatic image quality assessment with application in biometrics," in *Proceedings of the International Conference on Computer Vision and Pattern Recognition Workshop*, 2006. 76

[59] T. Gandhi and M. Trivedi, "Panoramic appearance map (pam) for multi-camera based person re-identification," in *Video and Signal Based Surveillance, 2006. AVSS '06. IEEE International Conference on*, nov. 2006, p. 78. 46

[60] N. Gheissari, T. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 1528–1535. 46

[61] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *10th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, 09/2007 2007. 46

[62] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proceedings of the 10th European Conference on Computer Vision 2008*, 2008. 46, 60, 65

[63] R. Gross, S. Baker, I. Matthews, J. Cohn, and T. Kanade, "Multi-pie," in *Proceedings of the Eighth IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2008. 95

[64] Grupo de Reconocimiento Facial e Vision Artificial, "FRAV2D face dataset," http://www.frav.es/databases/FRAV2d/. [Online]. Available: http://www.frav.es/databases/FRAV2d/ 95

[65] ——, "FRAV3D face dataset," http://www.frav.es/databases/FRAV3d/. [Online]. Available: http://www.frav.es/databases/FRAV3d/ 97, 101

[66] S. Gupta, K. R. Castleman, M. K. Markey, and A. C. Bovik, "Texas 3D face recognition database," in *Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation*, 2010, pp. 97–100. 96, 101

[67] R. Hartley, "Self-calibration from multiple views with a rotating camera," 1994, pp. 471–478. 16

[68] E. Hayman, T. Thorhallsson, and D. W. Murray., "Zoom-invariant tracking using points and lines in affine views - an application of the affine multifocal tensors," in *Proceedings of the IEEE International Conference on Computer Vision*, 1999. 17

[69] R. He, W.-S. Zheng, and B.-G. Hu, "Maximum correntropy criterion for robust face recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 8, pp. 1561 –1576, aug. 2011. 49, 109

[70] J. Heo and M. Savvides, "Gender and ethnicity specific generic elastic models from a single 2d image for novel 2d pose face synthesis and recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 12, pp. 2341–2350, Dec. 122

[71] M. Hirzer, P. Roth, and H. Bischof, "Person re-identification by efficient impostor-based metric learning," in *Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on*, Sept. 2012, pp. 203 –208. 60, 64, 65, 67

[72] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," ser. ECCV'12, 2012. 48, 60, 64, 65, 67, 70

[73] C. Huang and R. Nevatia, "High performance object detection by collaborative learning of joint ranking of granules features," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, june 2010, pp. 41 –48. 32

[74] J. Huang, V. Blanz, and B. Heisele, "Face recognition with support vector machines and 3d head models," in *Workshop on Pattern Recognition with Support Vector Machines*, 2002, pp. 334–341. 108

[75] M. Isard and A. Blake, "Condensation – conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, pp. 5–28, 1998. 32

[76] N. Jojic, A. Perina, M. Cristani, V. Murino, and B. Frey, "Stel component analysis: Modeling spatial correlations in image class structure," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, june 2009, pp. 2044 –2051. 47

[77] T. Kanade and A. Yamada, "Multi-subregion based probabilistic approach toward pose-invariant face recognition," in *Computational Intelligence in Robotics and Automation, 2003. Proceedings. 2003 IEEE International Symposium on*, vol. 2, 2003, pp. 954–959 vol.2. 121

[78] C. Kang, S. Liao, S. Xiang, and C. Pan, "Kernel sparse representation with local patterns for face recognition," in *ICIP*, 2011, pp. 3009–3012. 49, 109

[79] S. Katz, A. Tal, and R. Basri, "Direct visibility of point sets," in *ACM SIGGRAPH 2007 Papers*, ser. SIGGRAPH '07. ACM, 2007. 131

[80] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *International Symposium on Mixed and Augmented Reality*, 2007. 16, 22

[81] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Intern. Conf. on Computer Vision and Pattern Recognition*, 2012. 48

[82] E. H. L. De Agapito and I. Reid, "Self-calibration of rotating and zooming cameras," no. OUEL 0225/00, September 2000. 20

[83] B. Leibe, A. Leonardis, and B. Schiele., "Robust object detection with interleaved categorization and segmentation," *International Journal of Computer Vision Special Issue on Learning for Recognition and Recognition for Learning*, vol. 77, no. 1-3, pp. 259–289, 2008. 32

[84] A. Li, S. Shan, X. Chen, and W. Gao, "Maximizing intra-individual correlations for face recognition across pose differences," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 605–611. 120, 121, 122

[85] S. Li, X. Liu, X. Chai, H. Zhang, S. Lao, and S. Shan, "Morphable displacement field based image matching for face recognition across pose," in *Proceedings of the 12th European Conference on Computer Vision - Volume Part I*, ser. ECCV'12, 2012, pp. 102–115. 122, 135, 136, 138

[86] D. Liebowitz and A. Zisserman, "Metric rectification for perspective images of planes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1998. 26

[87] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *Proceedings of the International Conference on Image Processing*, 2002. 9, 77

[88] G. Little, S. Krishna, J. Black, and S. Panchanathan, "A methodology for evaluating robustness of face recognition algorithms with respect to variations in pose angle and illumination angle," in *International Conference on Acoustics, Speech, and Signal 2005*, vol. 2, 2005, pp. 89–92. 112

[89] Y. Liu, S. Ge, C. Li, and Z. You, "k-ns: A classifier by the distance to the nearest subspace," *Neural Networks, IEEE Transactions on*, vol. 22, no. 8, pp. 1256–1268, 2011. 62

[90] S. Lovegrove and A. J. Davison, "Real-time spherical mosaicing using whole image alignment," in *Proceedings of European Conference on Computer Vision*, 2010. 16, 17

[91] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal on Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004. 23

[92] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, Mar. 2010. 70, 111

[93] J. Marín, D. Vázquez, D. Gerónimo, and A. M. López, "Learning appearance in virtual scenarios for pedestrian detection." in *CVPR*, 2010, pp. 137–144. 108

[94] A. Mian, M. Bennamoun, and R. Owens, "An efficient multimodal 2d-3d hybrid approach to automatic face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 11, pp. 1927 –1943, nov. 2007. 124

[95] A. Moreno and y A.Sanchez, "Gavabdb: A 3D face database," in *Proceedings of the 2nd COST Workshop on Biometrics on the Internet: Fundamentals, Advances and Applications*, 2004, pp. 77–82. 97, 101

[96] K. Nasrollahi and T. B. Moeslund, "Complete face logs for video sequences using face quality measures," *IET Signal Processing*, vol. 3, no. 4, pp. 289–300, 2009. 76, 83, 85, 86

[97] K. Okuma, A. Taleghani, N. de Freitas, J. J. Little, and D. G. Lowe, "A boosted particle filter: Multitarget detection and tracking." in *Proceedings of the European Conference on Computer Vision*, 2004. 18, 31, 32, 39, 40

[98] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D face model for pose and illumination invariant face recognition," in *Proceedings of the 6th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS)*, 2009. 96, 101

[99] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, june 2005, pp. 947 – 954 vol. 1. 98, 101

[100] P. Phillips, W. Scruggs, A. O'Toole, P. Flynn, K. Bowyer, C. Schott, and M. Sharpe, "FRVT 2006 and ICE 2006 large-scale experimental results," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 5, pp. 831–846, 2010. 106

[101] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR, 2011, pp. 1201–1208. 31, 32, 37, 38, 39, 40

[102] U. Prabhu, J. Heo, and M. Savvides, "Unconstrained pose-invariant face recognition using 3d generic elastic models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 10, pp. 1952–1961, Oct. 122

[103] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," in *BMVC*, 2010, pp. 1–11. 48, 60, 64, 65

[104] D. Reid, "An algorithm for tracking multiple targets," *IEEE Transactions on Automatic Control*, vol. 24, no. 6, pp. 843 – 854, dec 1979. 32

[105] D. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, vol. 77, no. 1, pp. 125–141, May 2008. 81, 82

[106] J. Ruan and J. Yin, "Multi-pose face detection using facial features and adaboost algorithm," in *Proceedings of the Second International Workshop on Computer Science and Engineering*, 2009. 78

[107] X. W. Rui Zhao, Wanli Ouyang, "Unsupervised salience learning for person re-identification," in *CVPR*, 2013. 48, 60, 64, 65, 67

[108] A. Savran, N. Alyuz, H. Dibeklioglu, O. Celiktutan, B. Gokberk, B. Sankur, and L. Akarun, "Bosphorus database for 3D face analysis," in *Proceedings of the First COST 2101 Workshop on Biometrics and Identity Management (BIOID 2008)*, 2008. 97, 101

[109] W. R. Schwartz and L. S. Davis, "Learning discriminative appearance-based models using partial least squares," in *Proceedings of the 2009 XXII Brazilian Symposium on Computer Graphics and Image Processing*, ser. SIBGRAPI '09, 2009, pp. 322–329. 48, 67

[110] Y. Seo, S. Choi, H. Kim, and K.-S. Hong, "Where are the ball and players? soccer game analysis with color based tracking and image mosaick," in *Proceedings of the International Conference on Image Analysis and Processing*, 1997. 18

[111] A. Sharma, M. A. Haj, J. Choi, L. S. Davis, and D. W. Jacobs, "Robust pose invariant face recognition using coupled latent space discriminant analysis," *Comput. Vis. Image Underst.*, vol. 116, no. 11, pp. 1095–1110, Nov. 2012. 122, 138

[112] H.-Y. Shum and R. Szeliski, "Panoramic image mosaics," Tech. Rep., 1997. 16

[113] S. Sinha and M. Pollefeys., "Towards calibrating a pan-tilt-zoom cameras network," *P. Sturm, T. Svoboda, and S. Teller, editors, OMNIVIS*, 2004. 16, 17

[114] D. Song and K. Goldberg, "A minimum variance calibration algorithm for pan-tilt robotic cameras in natural environments," in *in IEEE International Conference on Robotics and Automation (ICRA*, 2006. 17

[115] S. Stalder, H. Grabner, and L. V. Gool, "Beyond semi-supervised tracking: Tracking should be as simple as detection, but not simpler than recognition," in *OLCV*, 2009. 80

[116] M. Subasic, S. Loncaric, T. Petkovic, H. Bogunovic, and V. Krivec, "Face image validation system," in *Proceedings of the International Symposium on Image and Signal Processing and Analysis*, 2005. 76

[117] H. Tao, H. S. Sawhney, and R. Kumar, "Object tracking with bayesian estimation of dynamic layer representations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 75–89, 2002. 18

[118] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society (Series B)*, vol. 58, pp. 267–288, 1996. 52

[119] C. Tomasi and T. Kanade, "Detection and tracking of point features," International Journal of Computer Vision, Tech. Rep., 1991. 32

[120] B. Tordoff and D. Murray, "Reactive control of zoom while fixating using perspective and affine cameras," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 1, pp. 98–112, 2004. 17, 18

[121] R. Y. Tsai and R. K. Lenz, "A new technique for fully autonomous and efficient 3d robotics hand-eye calibration," in *Proceedings of the 4th international symposium on Robotics Research*, 1988, pp. 287–297. 17

[122] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on riemannian manifolds," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 10, pp. 1713 –1727, oct. 2008. 27

[123] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 2001. 9, 77, 78, 88

[124] P. A. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004. 110

[125] S. G. W. Zheng and T. Xiang., "Transfer re-identification: From person to set-based verification." in *In Proc. IEEE Conference on Computer Vision and Pattern Recognition, Providence, Rhode Island, USA,*, ser. CVPR'12, June 2012. 48

[126] M. Weber, "Caltech frontal face dataset," http://www.vision.caltech.edu/. [Online]. Available: http://www.vision.caltech.edu/html-files/archive.html 95

[127] B. Williams, G. Klein, and I. Reid, "Real-time SLAM relocalisation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2007. 16

[128] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 2, pp. 210 –227, feb. 2009. 45, 49, 54, 63, 109, 111, 121

[129] Y. Wu, M. Minoh, M. Mukunoki, and S. Lao, "Set based discriminative ranking for recognition," in *ECCV*, 2012, pp. 497–510. 48, 60, 64, 65

[130] Z. Wu and R. Radke, "Keeping a pan-tilt-zoom camera calibrated," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, p. 1, 2012. 14, 17

[131] R. Xiao, M. Li, and H. Zhang, "Robust multipose face detection in images," *IEEE Transactions on Circuits System Video Technology*, vol. 14, no. 1, pp. 31–41, 2004. 78

[132] X. Yan, X. Wu, I. Kakadiaris, and S. Shah, "To track or to detect? an ensemble framework for optimal selection," in *ECCV*, 2012. 31, 32, 39, 40

[133] B. Yang and R. Nevatia, "Online learned discriminative part-based appearance models for multi-human tracking," in *ECCV*, 2012. 31, 32, 37, 38, 39, 40

[134] D. Yi, Z. Lei, and S. Li, "Towards pose robust face recognition," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013, pp. 3539–3545. 10, 120, 122, 129, 130, 133, 135, 136

[135] N. d. F. Yizheng Cai and J. Little., "Robust visual tracking for multiple targets." *Proceedings of the European Conference on Computer Vision*, 2006. 18

[136] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local gabor binary pattern histogram sequence (lgbphs): a novel non-statistical model for face representation and recognition," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 1, oct. 2005, pp. 786 – 791 Vol. 1. 135, 138

[137] X. Zhang and Y. Gao, "Face recognition across pose: A review," *Pattern Recogn.*, vol. 42, no. 11, pp. 2876–2896, Nov. 2009. 108

[138] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys*, vol. 35, pp. 399–458, December 2003. 94

[139] W. Zheng, S. Gong, and T. Xiang, "Re-identification by relative distance comparison," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PP, no. 99, p. 1, 2012. 48, 60, 65

[140] W.-S. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *BMVC*, 2009. 47, 60, 64, 65

[141] S. Zhou, R. Chellappa, and B. Moghaddam, "Visual tracking and recognition using appearance-based modeling in particle filters," 2003. [Online]. Available: citeseer.ist.psu.edu/zhou04visual.html 80

[142] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, june 2012, pp. 2879 –2886. 126, 130

[143] A. Zygmunt, P. Bródka, P. Kazienko, and J. Koźlak, "Key person analysis in social communities within the blogosphere," *j-jucs*, vol. 18, no. 4, pp. 577–597, feb 2012. 1