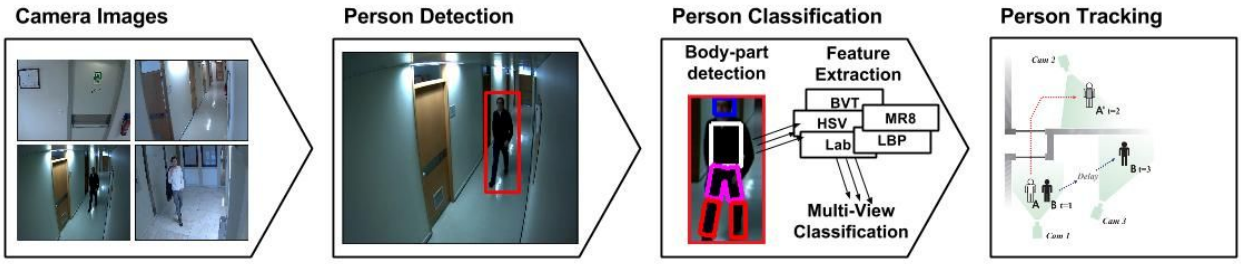
TÉCNICO LISBOA

UNIVERSIDADE DE LISBOA

INSTITUTO SUPERIOR TÉCNICO



Automatic Person Re-Identification

for Video Surveillance Applications

Dario António Bacellar Figueira

Supervisor: Doctor Alexandre José Malheiro Bernardino
Co-Supervisor: Doctor Jacinto Carlos Marques Peixoto do Nascimento

Thesis approved in public session to obtain the PhD Degree in
Electrical and Computer Engineering

Jury final classification: Pass with Distinction

Jury

Chairperson: Chairman of the IST Scientific Board

Members of the Committee:

Doctor Shaogang Gong
Doctor Jorge dos Santos Salvador Marques
Doctor Jaime dos Santos Cardoso
Doctor João Paulo Salgado Arriscado Costeira
Doctor Alexandre José Malheiro Bernardino
Doctor Jacinto Carlos Marques Peixoto do Nascimento

2016

UNIVERSIDADE DE LISBOA

INSTITUTO SUPERIOR TÉCNICO

Automatic Person Re-Identification

for Video Surveillance Applications

Dario António Bacellar Figueira

Supervisor: Doctor Alexandre José Malheiro Bernardino
Co-Supervisor: Doctor Jacinto Carlos Marques Peixoto do Nascimento

Thesis approved in public session to obtain the PhD Degree in
Electrical and Computer Engineering

Jury final classification: Pass with Distinction

Jury

Chairperson: Chairman of the IST Scientific Board

Members of the Committee:

Doctor Shaogang Gong, Professor, School of Electronic Engineering and Computer Science, Queen Mary University of London, UK

Doctor Jorge dos Santos Salvador Marques, Professor Associado (com Agregação) do Instituto Superior Técnico da Universidade de Lisboa

Doctor Jaime dos Santos Cardoso, Professor Associado da Faculdade de Engenharia da Universidade do Porto

Doctor João Paulo Salgado Arriscado Costeira, Professor Associado do Instituto Superior Técnico da Universidade de Lisboa

Doctor Alexandre José Malheiro Bernardino, Professor Associado do Instituto Superior Técnico da Universidade de Lisboa

Doctor Jacinto Carlos Marques Peixoto do Nascimento, Professor Auxiliar do Instituto Superior Técnico da Universidade de Lisboa

Funding Institutions
Fundação para a Ciência e Tecnologia

2016

ABSTRACT

Re-Identification is the problem of associating identities to detections of people over a network of cameras. Occlusions, changes in illumination conditions, different camera settings, view angles and pose, are visual contingencies that contribute to make re-identification a challenging problem in video-surveillance systems, specially in camera networks with non-overlapping fields of view. A practical re-identification system requires several components: person detection, feature extraction, classification and finally tracking across cameras. For the evaluation and deployment of the algorithms, suitable datasets, evaluation metrics and data presentation formats are needed.

In this work the re-identification problem is addressed in many perspectives. We propose novel methods for (i) dealing with failures and errors in detection; (ii) feature extraction using semantic body part segmentation; (iii) classification using Multi-View optimization techniques; (iv) temporal integration by window-based classifiers; (v) evaluation and data presentation for automated systems; (iv) and inter-camera tracking using a Multiple Hypothesis Tracker. The presented methodologies are evaluated in several datasets, including a novel high-definition dataset developed in-house, with applications to re-identification in camera networks.

With the aim of fully automating the re-identification procedure, it was proposed the integration of pedestrian detection methods with the classification stage of re-identification, and an evaluation of the issues arising from that integration was performed. In particular a false positive class was trained to tackle the false positives arising from the detection stage. For feature extraction, the effect of detecting and dividing the human body in semantically valid parts, such as dividing by the waist, or legs, torso and head, was evaluated. Extracting features from these local regions produces richer descriptors of person's appearance and increases recognition results consistently. For classification, a Multi-View semi-supervised optimization formulation was used, which integrates in a principled way several features (called views). The stated formulation allows for an optimal closed form solution which assures a fast learning. The semi-supervised aspect of the algorithm is well suited to the re-identification problem, where typically there are few labeled samples and a large number of unlabeled samples. To enhance performance of any single-frame classifier, a window-based wrapper for the classifier was proposed, that filters classification results according to the temporal coherence of pedestrian appearances. Finally, for inter-camera tracking the Multiple Hypothesis Tracker was used that keeps in memory multiple probable states of the world, which allows the tracker to update its belief based on both past and new information, being able to actually correct previous tracking association mistakes.

This work spans multiple facets of the video-surveillance problem, with a strong focus on autonomy and usability, thus strongly contributing towards

the wide applicability of re-identification systems in practical real-life scenarios.

Key-words: Re-Identification, Pedestrian Detection, Camera Networks, Video Surveillance, Inter-camera Tracking

RESUMO

Re-identificação consiste em fazer seguimento das pessoas entre cameras. É um problema ainda em aberto devido à grande variabilidade da aparencia das pessoas nas imagens de diferentes cameras (e até na mesma camera). Oclusões, diferenças de iluminação, diferenças na pose, diferenças no balanço das cores de cada camera, diferenças no ângulo de visionamento da camera e às vezes mudança de roupa das pessoas, são tudo coisas que dificultam a re-identificação.

É um problema interessante pois o número sempre crescente de cameras de video-vigilância existentes hoje já ultrapassa a capacidade de monitorização dos seguranças humanos. Não só é uma aplicação necessessária na segurança, mas também potencia todo um leque de outras aplicações tais como espaços inteligentes, video-jogos, pesquisa sobre as actividades das pessoas no dia-a-dia.

Neste trabalho abordam-se todos os estágios da re-identificação, desde a detecção de pedestres, passando pela classificação dos mesmos, e finalmente fazendo seguimento entre cameras. Propõe-se um método de extração de caracteristicas locais das pessoas baseado na detecção das partes do corpo. Confirma-se que a extração local de características aumenta a performance da re-identificação. Utiliza-se um classificador semi-supervisionado chamado Multi-View para aproveitar o grande número de imagens não identificadas que existe neste meio. Explora-se a coerência temporal das pessoas nas imagens de video para aumentar a performance. Propõem-se estratégias para lidar com os problemas que advêm de se ter detecção automática de pedestres. Tais como um filtro de detecções parciais e uma classe para a classificação de falsos positivos. Propõe-se também metricas de avaliação do sistema integrado para correctamente medir o impacto das falhas de detecção que não são consideradas no estado-da-arte da re-identificação. Por fim apresenta-se os resultados de uma forma inovadora que poupa no tempo de visionamento do utilizador.

Testou-se os variados algoritmos em várias bases de dados de imagens. Com este trabalho, de aplicação geral, espera-se que a re-identificação se torne uma realidade prática num futuro próximo.

Palavras-chave: Re-Identificação, Detecção de Pedestres, Rede de Câmeras, Video Vigilância, Seguimento entre Câmeras

There is an uncertainty relationship between truth and clarity.

— Niels Bohr

ACKNOWLEDGMENTS

A big heartfelt thank you to all who supported me and made this thesis possible, thank you.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

BB Bounding Box

BVT Black-Value-Tint histogram

CMC Cumulative Matching Characteristic curve

FP False Positive

GT Ground Truth

HSV Hue-Saturation-Value histogram

Lab Lightness color-opponent histogram

LBP Local Binary Patterns

MAP Maximum a Posteriori

MHT Multiple Hypothesis Tracking

MR8 Maximum Response Filter Bank

MSCR Maximally Stable Color Regions

MD Missed Detection

NN Nearest-Neighbor

NT New Target

PD Pedestrian Detection

PS Pictorial Structures

RE-ID Re-Identification

RGB Red Green and Blue color model

RKHS Reproducing Kernel Hilbert Space

SIFT Scale Invariant Feature Transform

SURF Speeded-Up Robust Features

1 INTRODUCTION

In this work the problem of re-identification of people in camera networks is addressed (see Figure 1). Given a set of pictures of previously identified persons, a practical RE-ID system must locate and recognize such people in the stream of images that flows from a camera network, past or present.

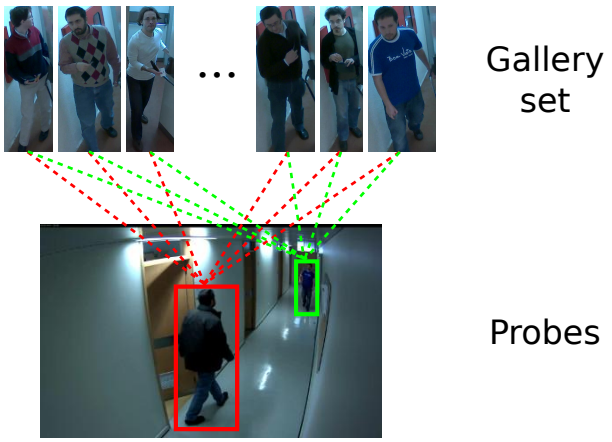


Figure 1: A typical re-identification algorithm is based on a gallery set: a database that contains the persons to be re-identified at evaluation time. People detected in other images (probes) are matched to such database with the intent of recognizing their identities. Classically, re-identification algorithms are evaluated with manually cropped probes. In this work, instead, we study the effect of using an automatic pedestrian detector to propose probes.

Most works define re-identification as matching pedestrian images only from different cameras with non-overlapping fields of view [19, 22, 74, 77, 26, 132, 85, 122]. Some works define re-identification allowing the matching to happen also for images in the same camera [91, 69, 49, 108]. The problem also has many manifestations in other application domains. For instance:

- In the field of tracking, a similar problem is known as "re-acquisition" [76] when the aim is to associate a target (person) when it is temporarily occluded during the tracking in a single camera view. However, in tracking the association is of targets in contiguous time and space, and the more general re-identification can have the targets separated by large time scales and image positions;
- In a human–robot interaction scenario, solving the re-identification problem can be considered as "non-cooperative target recognition" [70], where the identity of the interlocutor is to be maintained, allowing the robot to be continuously aware of the surrounding people;

- In larger distributed spaces such as airport terminals and shopping malls, re-identification is mostly considered as the task of "object association" [54, 88] in a distributed multi-camera network, where the goal is to keep track of an individual across different cameras with non-overlapping field of views.

In this work, re-identification is defined as undeniably linked to an identification phase where the association of some images to identifying labels was done through some proper high-confidence method, such as strong biometrics (*e.g.,* fingerprint, retinal scan, face recognition) or through the presentation of an unique ID card, *i.e.,* at a controlled entrance where those initial images could be taken, or even by manually labeled after human inspection. These identified images form a gallery, that are the basis against which the non-labeled images called probes are re-identified.

Video surveillance cameras are now ubiquitous in most malls and in some city streets as well (*i.e.,* over half a million cameras in London [93]). Typically, these images are inspected by human operators to detect abnormal events in the video streams. The classical application for RE-ID is then video surveillance for security in large commercial spaces like shopping centers or office buildings. Other applications of RE-ID lay in smart spaces, such as intelligent office buildings, which require the detection and identification of its occupants in order to control the environment intelligently, *e.g.,* change the background music, illumination style and temperature given the rooms' occupants' preferences. Re-identification algorithms also enable tracking systems to link person's trajectories across multiple cameras in a network. This ability is essential to support research in several other fields, *e.g.,* modeling activities, mining physical social networks and human-robot interaction.

Most of the practical applications of a RE-ID system can be formulated by the following three queries to the system (also listed in Table 1):

- "$\mathcal{Q}_1$: *Who is X?"* In this query the input is a bounding box (cropped image) containing an unknown person (a probe sample X), and the output should be the ID of the person as stored in the gallery;
- "$\mathcal{Q}_2$: *Where is John?"* In this query the input consists of the ID of the desired person (John), video sequences, possibly from multiple cameras, and the output are sub-sequences of the video-clips containing John;
- "$\mathcal{Q}_3$: *Where else is X?"* In this query the input is a bounding box containing a person, video sequences, possibly from multiple cameras, while the output are sub-sequences of the video-clips containing the person X.

To tackle the mentioned applications one possible architecture for re-identification is illustrated in Figure 2. It is composed of several stages:

	Input	Output
$\mathcal{Q}_1$: Who is X (test sample)	Bounding Box	Person ID
$\mathcal{Q}_2$: Where is John (train sample)	John ID, Video Sequences	Video-clips with John
$\mathcal{Q}_3$: Where else is X (test sample)	Bounding Box, Video Sequences	Video-clips with X

Table 1: Typical Re-Identification questions to address: ($\mathcal{Q}_1$) Given a bounding box containing a person output its ID; ($\mathcal{Q}_2$) Given a person ID, and some video sequences where to search, output video-clips containing images of that person; ($\mathcal{Q}_3$) Given a bounding box containing a person, and some video sequences where to look, output video-clips containing images of that same person.

- People must be identified at some point before re-identification can be enacted, either by some hard biometric sensor or some uniquely identifying card. Thus images of people are associated with their identifications creating a gallery of identified people (upper block in Figure 2).

1. People must be extracted from the camera views (in Figure 2b), either by manual or automatic means;
2. Distinctive features need to be extracted from the individuals to discriminate between them (in Figure 2c);
3. Individual detections are then matched against the gallery (also in Figure 2c); and finally
4. Re-identified individuals can be tracked over the camera network (in Figure 2d).

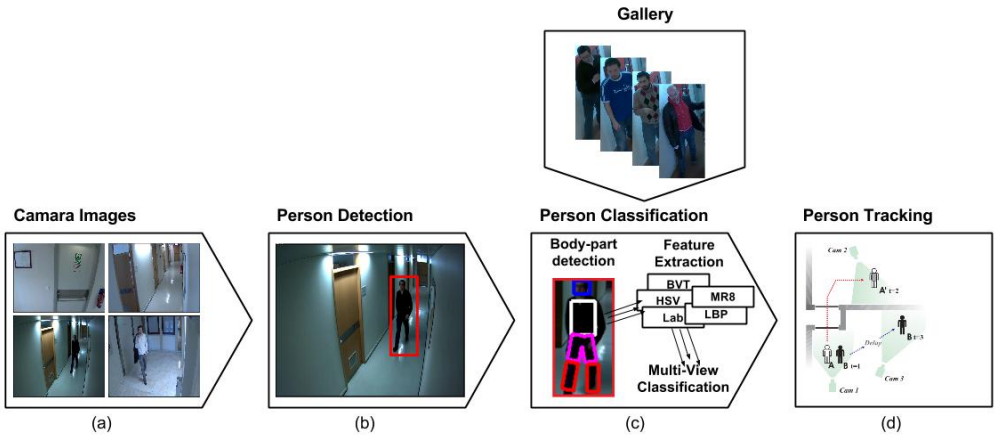


Figure 2: A possible general architecture of the automated Re-Identification problem is presented here. It presents the major components used in a re-identification system.

1.1 CHALLENGES OF RE-ID

Re-Identification is a challenging problem with several inherent difficulties. A wide range of people's body motion and poses, self-occlusions, occlusions by others, and possibly even changing clothes make the recognition problem already quite challenging by itself (see Figure 3). When the different opto-electric characteristics of distinct cameras and the different possible viewing angles and distances are taken into consideration, re-identification becomes even more challenging. In fact, all this is known to cause images of the same person to sometimes be more different than images from two separate people.

Figure 3: The problems of different camera color balance, different illumination, different camera angle, different pedestrian pose and different person attire are illustrated in this figure. Between the two figures in the left one can observe different illumination, color balance and pedestrian pose. Between the right figure and the other two, it can be observed different clothes in the designated pedestrian.

Another problem is that some features may not be usable in the whole camera network, or even in certain locations of a camera view. Ideally, a hard biometric feature such as face recognition would be the principal feature used in re-identification. It has a high degree of reliability, but requires close-up images, frontal views and some user collaboration. Therefore, it cannot be used in most common scenarios. Different features then have to be used in different places, given the different sensors available and different geometry of the view. In most locations today only cameras with moderate resolution are available, which limits the features that may be used to mostly color and texture. Even motion is hard to use with uncooperative people in unconstrained environments. If higher definition cameras are available, or the geometry of the space allows the cameras to be closer to the faces of the subjects, face recognition may be employed, and such hard biometrics can then lend confidence to the "soft" biometrics of clothing color and texture that are used in the rest of the cameras in the network.

A real system also requires automatic detection of the pedestrians which leads to a host of issues such as false positive detections, unreliable bounding boxes, and missed pedestrians – all issues that further hinder RE-ID. Another issue that follows from the existence of false positives is the lack of confidence

of the users in the system if it generates many false alarms. But when the system is tuned to be more discretionary, rejecting detections that are considered to be false alarms, some true detections will be discarded as well, leading to an increase in missed detections. A re-identification system will have to walk the fine line between false alarms and missed detections. Too many false alarms or too many missed detections will lead to lack of confidence and thus rejection of the system by the users. Another part of the system that is challenging to automate, is the creation and maintenance of the gallery. At the current time, real-world systems mostly rely on human intervention to do or at least verify changes to the gallery, which guarantees a strong identification stage. An automated system will have to manage what was previous manual human intervention. If in a office-space building, the system may have the possibility of strong identification (by biometrics or identification card), and then the issue of adding new people to the gallery can be trivial. In an open-space like a shopping center, a re-identification system can help a human by automatically re-identifying persons in a gallery. However, since no real possibility exists for strong identification, weaker methods must be used to determine if each detection belongs to an existing person in the gallery, or if this is a new subject to be added. This would be an enhancement over the current human-managed systems in open-spaces, since at present, maintaining a small gallery of persons of interest (*e.g.,* known thieves) that is updated very slowly (*e.g.,* when new thefts happen) is what is currently available.

Another challenge lies in the interaction with the user, how to present results to the user, since human attention span is limited. Typically, re-identification has been performed by human operators, that inspect the video feeds to detect abnormal events and persons of interest. However, human attention is limited and most often the human operators can not cope with the huge amount of available information and many abnormal events and persons of interest may pass unnoticed. This means the full potential of surveillance systems today is still under-explored. Hence, the use of automatic person re-identification methods to aid human operators focus their attention on targets of interest is *essential*. The simplest way is to present all the frames an individual was re-identified in. A more sophisticated approach could be collating the contiguous frames in time, and presenting short videos instead. If the topology of the camera placements is available to the system, it can draw a probable path an individual took, given the temporal constraints and the re-identifications in each camera.

The evaluation and performance characterization of RE-ID systems requires the creation of labeled datasets. Creating a dataset with images that contains the diversity of situation that may arise in a practical scenario is an issue in itself, but they are of the utmost importance to properly benchmark algorithms and drive the state-of-the-art to higher levels. A dataset needs to be challenging to properly test the algorithms and highlight their flaws and bottlenecks. But, capturing video feeds in scenarios of interest and then manually labeling all persons appearances in the captured images is a costly process. Further-

more, carefully chosen benchmark metrics are required to properly compare among different algorithms, specially when automatic detection is used.

1.2 THIS THESIS

In this thesis, I address some of the current challenges in RE-ID systems and contribute with novel approaches and algorithms beyond the state-of-the-art.

- The person appearance variability was approached, first by improving the feature extraction process with more relevant local features, and second, by applying a state-of-the-art semi-supervised classification algorithm to the problem of re-identification (see Figure 2c). Finally, by an over-arching tracking algorithm that is able to correct past re-identifications and thus ameliorate the issue of people changing clothes (see Figure 2d).

- System automation was approached by defining an architecture for automated RE-ID. This architecture includes a pedestrian detection algorithm to automate the detection part of the re-identification pipeline (Figure 2b). However, such automation also introduces errors (unreliable bounding boxes, false and missed detections) that were approached in three ways: first, by a time-filter wrapper for the classifier that eliminates spurious false positives and re-captures some missed detections; second, by the development of a module tailored to address the remaining false positives; and finally, by using a local feature extraction method that ameliorates the issue of unreliable bounding boxes. Most of the work in the literature assumes perfect detections and thus is unable to cope with these issues.

- The issue of system output to the user was approached by presenting the results as video clips instead of still images, so that the load to the limited human span is reduced.

- The issue of algorithm evaluation was approached in two different ways. First, by participating in the development of one of the most complete and challenging dataset for re-identification to date. The dataset is larger than most, and contains examples of many of the above mentioned issues, such as many examples of high pedestrian appearance variability due to different poses, viewing angles, and opto-electric camera characteristics and even cloth changes. Second, more informative metrics were applied to complement the standard metric used in most of the literature of re-identification. The standard metric does not highlight the influence of false positives and missed detections, that must be taken into account in automated systems.

1.3 STATE-OF-THE-ART AND A TAXONOMY OF RE-IDENTIFICATION SYSTEMS

Current research activity on re-identification is mainly focused in two areas: (a) the development of feature representation which properly discriminates the identities of people, either by manual design or learning from the data [57**?** , 73]; and (b) the development of matching methods [104, 131]. However, to take RE-ID towards practical applications there are many other dimensions of interest, such as generalization to different scenarios, the way data is presented to the user, the time span of applications, among others.

Following an extensive review of the literature, a more complete taxonomy of the state-of-the-art was defined (see Table 2).

The main dimensions identified are described below:

Open vs Closed spaces: One of the dimensions of the taxonomy is how persons are introduced on the gallery. If the space is closed, there exists a controlled entry where good quality images can be taken and the identity of each person is verified. Thus the gallery is created at an identification stage, prior to re-identification, without any uncertainty in the person identity. For open scenarios there is no controlled entry and the difficulty level rises, since any number of new pedestrians may cross the system field of view and thus the gallery must be maintained dynamically, adding or deleting new entries as seems necessary at run-time. Errors in the gallery maintenance will be another issue that will require attention. Only a few works have tackled the open-space scenario, such as the pioneering work of Gong *et al* [88], where the time delayed correlation between camera events is used to aid the matching between two detections in different cameras. In the present thesis a closed space scenario is considered.

Manual vs Automatic probe: Another dimension is the way probe images are selected. Most algorithms in the literature assume it is a human operator that draws a bounding box around a person in an image to query the system for other instances of the same person, but an automatic methodology is necessary for many real applications, for instance tracking. In the automatic case, problems like detection failures (false positives or missed detections), bounding box misalignment's and partial occlusions must be taken into consideration. Very few works approach the automatic probe case. This thesis proposes ways to tackle some of the issues that arise from detection failures [115, 48].

Single-shot vs Multi-shot: Another dimension that categorizes RE-ID algorithms is the use of an in-camera tracker, in which case the query data consists of more than one image per exemplar (multi-shot), otherwise only one image will be available per sample (single-shot). This determines if the application is single-shot or multi-shot. In this thesis results for the single-shot case are presented. The theory and presented methods can be readily extended to the multi-shot case. Note that, to the best of the author's knowledge, works that do **Multi-Shot** don't actually use an in-camera tracker, but just assume the human operator selects a sequence of contiguous bounding boxes of the

same person in a time window (or if dealing with a pre-recorded dataset, the manual annotations provided by the dataset).

Short-term vs Long-term: This dimension represents the time scale in which RE-ID is computed, that is, the temporal validity of a match between gallery and probe data. The methods presented in this thesis, as in most of related work in the literature, are only able to tackle the short-term case (typically under one day) because clothing color is the feature of election in the state-of-the-art approaches, and clothing constancy can be expected only in the short term. Note that except for the pioneering work of Nakajima *et al.* [98], that purposely runs a RE-ID experiment over the course of a few days, up to my knowledge no published work has tried to tackle the **Long Term** time scale.

Frame-based vs Video-based output: The type of output resulting from a query (*e.g.*, the queries in Table 1), can be composed of single frames or video-clips, and this forms another dimension of the problem. When the output must be checked by a human operator, less time would be spent analyzing a short video (multiple frames) than analyzing each of the individual frames. Therefore, a video-based output is proposed as a means to reduce the operator overload. To the best of my knowledge, the existing works in the literature just tackle the problem of determining the identity of the probe data, and so don't even provide frame output, only a classification for each probe sample.

This dimension of the taxonomy is not standard in the state-of-the-art since it is related to user interaction issues, that are novel contributions presented in this thesis, such as the type of output provided to a user. This contribution brings RE-ID systems closer to actual applications in the real-world.

	Probe		Scenario		Exemplar size		Time scale		Output	
	Manual	**Automatic**	**Open**	**Closed**	**Single**	**Multi**	**Short**	**Long**	**Frame**	**Video**
[100, 52, 57, 58, 127, 9, 104, 44] [6, 126, 34, 131, 82, 73, 72, 8] [78, 94, 83, 3, 46, 86, 116]	✓	×	×	✓	✓	×	✓	×	✓	×
[60**?** , 25, 17, 18, 11, 90] [129, 102, 120]	✓	×	×	✓	✓	✓	✓	×	✓	×
[121, 61, 118, 117, 111, 68, 10] [62, 12]	✓	×	×	✓	×	✓	✓	×	✓	×
Nakajima *et al.* [98]	×	✓...	×	✓	×	✓	✓	✓	✓	×
Gilbert *et al.* [54]	×	✓...	✓	×	×	✓	✓	×	✓	×
Gong *et al.* [88]	×	✓...	✓	×	✓	×	✓	×	✓	×
Bak *et al.* [29]	×	✓...	×	✓	×	✓	✓	×	✓	×
[67, 95]	×	✓...	✓	×	×	✓	✓	×	✓	×
Dario Figueira *et al.* [115, 47]	✓	✓	×	✓	✓	×	✓	×	✓	×
Dario Figueira *et al.* [48]	✓	✓	×	✓	✓	×	✓	×	✓	✓

Table 2: A taxonomy of the state-of-the-art in RE-ID (see text for the definitions). Note that all the works of other authors under the **Automatic** Probe generation only actually do semi-automatic, using an automatic algorithm to detect pedestrians which will have unreliable bounding-boxes but then manually removing all false positives and not dealing with the missed detections. Also note that many works only tackle the problem of determining the identity of the probe data (instead of providing frame output), and they have been classified in the **Frame Output** class since conceptually they can only show the user the frames in the gallery database.

Besides a couple of pioneering works [53, 88], that do correlation analysis between camera events, and attempt to associate person apparitions in different cameras together (instead of attempting to re-identify against a gallery) most of the standard RE-ID methodologies in the literature only focus on answering Q_1 (see Table 1) and work under the assumption of a closed space scenario (where the input bounding box is manually selected and the selected person is present in the gallery). They work with manual annotations of a "short" time scale (both under single and multi-shot approaches) as can be seen in the first three lines of Table 2. The aim of this work is taking RE-ID systems to novel application levels, where question Q_2 and Q_3 (see Table 1) are of practical relevance. By providing video-based output, instead of individual frames, it becomes easier for the system operator to perform queries Q_2 and Q_3 and obtain relevant results.

The work in this thesis, as indicated by the last couple of lines of Table 2, encompasses the following dimensions of the proposed taxonomy: (i) both manual annotation and automatic pedestrian detection for probe generation, (ii) closed scenario, (iii) single-shot, without a in-camera tracker, (iv) within the time-frame of a single day, and (v) providing both frame based output and video-clip based output.

In this section the RE-ID problem was classified in several dimensions and the works in the state-of-the-art were categorized into those dimensions. These works will be mentioned again in the next chapters as relevant related work for each of the proposed methods in this thesis.

1.4 CONTRIBUTIONS

In this work the problem of RE-ID is analyzed in several aspects. The work contributes towards the automation of RE-ID systems through the integration with PD algorithms. This integration must take into account the sources of errors still present in current pedestrian detection systems (see contribution 1 below). Contributions to deal with high pedestrian variability are proposed by enhancing the state-of-the-art in feature extraction and classification (see contributions 2, 3 and 4 below). To alleviate the cognitive load of the RE-ID operator in a practical surveillance system output is provided in the form of small video-clips (see contribution 5 below). Finally, contributions to algorithm evaluation were made through the development of a cutting edge dataset, and the application of complimentary performance metrics (see contribution 6). More concretely, the contributions are enumerated as follows:

1. Several problems arise when integrating PD with RE-ID. First, bounding boxes detected by automatic methods are often misaligned with the persons boundaries. By using body-part detectors [4] on the detection windows this problem is alleviated (Section 3.1.1). Second, state-of-the art automatic detection methods still produce frequent false positive detections. By training a class for the typical false positives in a certain

environment (Section 3.1.3) RE-ID quality can be significantly improved (work developed in [115, 48]).

2. To tackle the high variability of human appearance, body-part detection is also used in Section 3.2 (work developed in [44]). Body-part detection is applied to the human bounding boxes for local and more relevant feature extraction (an extension of the works [4, 25]). Bounding-boxes are thus divided in body parts, to be able to extract features from semantically meaningful local regions. This obviates the need for background subtraction and comparative analysis show that it improves results consistently with respect to many works in the state-of-the-art [104, 130, 73].

3. Also to tackle the high variability of human appearance, a state-of-the-art classifier [105] was applied to re-identification, described in Section 3.3.1 (developed in [46]). A semi-supervised Multi-View classification algorithm is used to take advantage of all the unlabeled test data and combine all extracted features. It has an optimal closed form solution that allows for fast learning. It copes well with small number of training samples, and the semi-supervised aspect of it is well suited to the re-identification problem, where it is common to have small training sets and large number of unlabeled samples. Results in Chapter 4 ground our assertion that this helps tackle pedestrian appearance variability.

4. To further enhance classification, in Section 3.3.2 a window-based classifier was proposed (developed in [48]). It exploits the temporal coherence of pedestrian appearances in each camera view, eliminating spurious mis-classifications by filtering the output from any single-frame classifier. Some missed detections of the detector are also recaptured, when those missed detections fall between correct re-identifications.

5. The window-based classifier also naturally suggests that output be provided in the form of video-clips, which alleviate the cognitive load of users that will review/validate the output (see Section 3.3.3 and [48]). This is the case since evaluating a small video-clip of a person's detections and re-identifications is much faster than doing the same for each individual frame.

6. Another point of contribution, to deal with the challenge of algorithm evaluation, was the participation in the development of one of the best datasets for evaluation of re-identification algorithms (see [116, 47]). This dataset contains many examples of the issues enumerated above, such as high pedestrian appearance variability from multiple poses, viewing angles, occlusions, opto-electric camera characteristics, and even changing clothes. Additionally, as described in Section 4.4.1, this work used metrics that complement the evaluation of RE-ID systems when they are integrated with a PD algorithm (work also developed in [48]). Metrics are proposed that assess the impact of false positives and missed

detections in the overall system, and that complement the usual metric employed by the RE-ID community (CMC curves).

1.5 WORK STRUCTURE

Chapter 3 reviews the background and related work relevant to this thesis, and Chapter 3 goes over the main work of this thesis:

1. in Section 3.1, an architecture for the integration of PD with RE-ID is presented and methods to address the issues that arise from that integration are discussed;

2. in Section 3.2, the problem of extracting features for re-identification from persons' bounding boxes is addressed. It is proposed the development of a semantic division of a pedestrian from where to extract descriptive features;

3. in Section 3.3.1 it is proposed the use of a semi-supervised formulation for classification, for RE-ID. The proposed method successfully fuses any number of different features (Multi-View), and copes well with a small number of training samples;

4. the enhancement of classification through the exploitation of the pedestrians' temporal coherence in described in Section 3.3.2 that details the window-based classifier;

5. the novelty of providing output as video-clips, to alleviate te users' cognitive load is presented in Section 3.3.3;

6. and finally, in Section 4.4.1 a novel metric is proposed to properly assess the weight of false positives and missed detections in RE-ID.

In Chapter 4 all the results from the comparative work done over the years is gathered. Finally in Chapter 5, conclusions are drawn, the possible future work is discussed and a list of the published works is presented.

Finally, Appendix A describes a labeling software that was used and improved on while helping in the creation of the HDA dataset [116]. Appendix B describes in some detail work on metric learning developed during the early years of this work but not central to the discussion.

2

BACKGROUND AND RELATED WORK

2.1 TYPICAL ARCHITECTURE FOR RE-ID

In this chapter the RE-ID problem is described in an overall perspective, taking into account all the modules and functionality required for a high degree of autonomy in video surveillance systems. One possible overall architecture of an automated re-identification problem is presented in Figure 2. Every RE-ID system is based on a gallery set and a probe. A gallery set is composed by images or sequences of images from a person to be recognized across the cameras of the network. The probe is an image of a person to be re-identified against the gallery images.

A gallery set is either acquired off-line or online. In the off-line version people are registered to be allowed to enter the space. In the online version the gallery is updated as people enter and exit the system. In the online version we can also distinguish between closed-spaces, where the gallery examples are acquired at special access points in the camera network, and open-spaces, where the gallery examples are acquired at any point.

Concerning the detection stage (Figure 2b), at runtime, persons are detected from the camera network's images. Detections are usually represented as bounding-boxes around the persons' images and can be obtained either by the system's operator manual intervention or automatically, by pedestrian detection algorithms or background subtraction methods. The process of RE-ID then consists in associating runtime person detections to the gallery examples. Analysis can be done at individual frames (single-shot) or with multiple frames from tracks within the same camera (multi-shot). Analysis is typically performed looking at several features extracted from the persons bounding boxes, *e.g.*, color, shape, texture, or motion. These features are then associated to examples in the gallery through appropriate classifiers. Classifiers range from as simple as NN to more complex supervised or semi-supervised methods. Finally the classified pedestrians are tracked throughout the camera network exploiting as much as possible the constraints in the network topology and human motions, *e.g.*, via Multiple Hypothesis Tracking (MHT).

In the following section are described in more detail the most important components necessary for real-world applications, which encompass some of the challenges tackled in this thesis.

2.2 COMPONENTS FOR RE-ID

2.2.1 *Pedestrian Detection*

Previous to enacting re-identification, pedestrians must be detected. Pedestrian Detection is a subject that has drawn much interest and is rich in the literature. The work available ranges from part-based detectors, which explicitly model the articulation of the human body (see [43, 103]), to monolithic detectors (see [31, 39, 20]), which associate one descriptor to one detection window.

In the beginning of this work a target detection algorithm by Boult *et al.* [21] was used to detect each pedestrian. But, given the difficulty of the pedestrian detection problem, and the desire to use "clean" (perfect) detections, manually annotated people detections was later used for the majority of this work. However, in the last stage of this thesis, an automatic pedestrian detection algorithm [114] was used, to, as already mentioned, study the effects of using not-perfect detections as input to the RE-ID algorithms. The issues were evaluated and some solutions to circumvent the errors were proposed [115, 48] (further details in Section 3.1).

At this point an in-camera tracker can be used to associate detections previous to the classification. If one does so, many images will be available per exemplar at the classification stage (multi-shot situation) allowing for more features to be extracted, averaging out noise, or even automatically picking images that seem to be the cleanest [120]. If no in-camera tracking is used, only one image per exemplar is available (single-shot situation).

2.2.2 *Features*

To perform re-identification it is needed to have selective and consistent features to be able to reliably distinguish different persons in a systematic way. The issue of manually designing the features or learning from the data which features are distinctive arises at this point. This is richly addressed in the literature (*e.g.*, [9, 2, 17]). One can manually design/choose a feature (*e.g.*, HSV histograms) to be used, or have several feature channels and combine/weight them in some fashion [42, 25], or attempt to determine for each test sample which features best describes it (*e.g.*, texture feature for a patterned shirt wearing person) and then use that feature in the classification stage [81].

The well know color features Hue-Saturation-Value histogram (HSV) and Lightness color-opponent histogram (Lab) have been the most widely used features in recent work. This happens because the majority of works address short term RE-ID scenarios where clothes color distribution is an important feature. HSV's color space is a more intuitive way to describe color than Red Green and Blue color model (RGB). Lab's color space was developed to be more perceptually relevant since small changes in human color perception match small changes in the Lab color space. Texture features like Local Binary Patterns (LBP) [2], Maximum Response Filter Bank (MR8) [75, 109], Gabor fil-

ters and Schmidt filters have also been used in re-identification [104], since for some pedestrians with textured appearance texture features are more appropriate than only color. LBP describes texture by means of patterns of relative brightness of pixels surrounding a central pixel. Gabor filters[50] were designed to detect edges. They are the result of a complex exponential modulated by a Gaussian window and subjected to scaling and rotation. Schmidt filters[110] are rotational invariant filters, designed to detect local *maxima* and *minima* of brightness. MR8 collects a set of 36 "Gabor-like" texture filters taking the maximum over many of them in such a fashion that MR8 is reduced to 6 edge detection texture filters invariant to rotation, plus 2 Schmidt-like filters – one for detecting *maxima* and one for detecting *minima* of brightness. Kovalev *et al.* [71] proposed the color co-occurrence correlogram. Hamboun *et al.* [60] applied Speeded-Up Robust Features (SURF) to re-identification, and Teixeira *et al.* [117] chose Scale Invariant Feature Transform (SIFT) to tackle this matching problem. A review of the literature provides the impression that texture is important for re-identification, but color still does the brunt of the work when discrimination people's appearance.

Of special note is Liu *et al.* [81] and Layne *et al.* [72]. Liu *et al.* employed a dynamic feature selection, using the feature type that works best for each kind of pedestrian clothes. The training set images are clustered based on the feature type (color or texture) that best discriminates among exemplars. For instance, the cluster of a texture feature contains people with very textured clothes like checkered shirts and the cluster of a color feature contains people with brightly colored clothes. Then at run-time, the incoming test image is mapped to the closest cluster, and the image is described with the corresponding feature. In that work only one feature vector per cluster is used, containing the feature that work best for each cluster. Layne *et al.* trained semantically human understandable attributes that are quite transferable across datasets. While the absolute performance gains from these initial works are not tremendous, the generalization properties they display are of great interest.

2.2.3 *Feature Extraction*

Not only *which* features to extract are important, but also *from where* in a detection bounding box is an issue up for research. The simplest way is to extract uniformly from the whole person image. When one realizes most images are of upright people (denominated pedestrians) and that their appearance varies most in the vertical dimension, a step further can be taken by dividing the pedestrian in horizontal stripes and extracting features accordingly [104].

For feature extraction, other works can also be adopted, such as Feltzenswalb [55] body-part detectors, or Pictorial Structures (PS) body part detectors [4]. Other works that don't extract features from semantically valid image regions usually divide the person detection bounding box in six horizontal stripes and extract features accordingly [130, 104]. This does not provide the best results

compared to body-part detection, but is still useful when comparing different classification algorithms.

2.2.4 *Classification*

Once features have been extracted from the image and descriptors of the detect people created, classification must be performed to associate the detection to the persons' information contained in the gallery. This is a rich field in the literature where we find many works using NN by direct distance minimization [79, 125, 100], while others use SVM or SVM-like approaches [8, 119, 104], and many other different classifiers abound in the literature.

Some works have empirically integrated different feature types using weighted average to join the output of different features [42, 25]. Others have concatenated several feature types [131, 81, 129], and relied on the classifier to exploit the information present on the different features.

The classification stage of RE-ID is rich with alternatives in the literature, either through learning or by simple direct distance minimization ([56, 79, 106, 9, 100, 111]). Distinct from his peers, Zheng *et al.* [131] learns a relative distance metric. Instead on focusing on minimizing the intra-class distance and maximizing the inter-class distances, he computes a metric such that, given triplets of images containing two images of one person and an image of a different one, the distance between images of different people is greater than the distance between images of a same person.

Still in the classification stage of re-identification, of special note is Tamar's work [8], that trains a SVM binary classifier to distinguish positive pairs (two concatenated feature vectors from a pair of images of a same person) from negative pairs (concatenated feature vectors of images pairs of different people), with comparable results to the state of the art. Although the performance is not significantly better than the other works in the state of the art, the method warrants notice for being such a simple and successful application of a known classifier.

In this thesis a classification approach that trains one classifier per feature (called "view") was applied [46]. This approach exploits the fact that some features outperform the others in some parts of the training data. It uses this higher performance to improve the re-identification performance of the other feature types in those same parts of the data. It is also a semi-supervised technique, allowing the exploitation of the many unlabeled data usually present in re-identification. In the next chapter it is demonstrated that this strategy achieves higher classification performances than many state of the art classifiers.

It is worth noting that given the high difficulty of the RE-ID problem in general, the classifier algorithms rarely give binary hard classifications, but instead output ranked lists or probability values for each gallery entry for a given probe sample. This can later be exploited when applying temporal filters and inter-camera tracking.

2.2.5 *Tracking*

Finally after pedestrians have been detected and classified the final goal can be accomplished: tracking people across the camera network.

At this level the topology of the network can be exploited. The map of the network can be manually defined or learnt automatically [53]) to allow for the exploitation of temporal constraints of pedestrian apparitions in different cameras. Because a pedestrian can't be in two cameras at the same time, the tracking stage will disregard or even correct some mistakes of the classification stage. Combining the tracking and classification stages will improve the overall performance. However, the errors from the tracking stage must also be deal with.

One baseline in tracking across multiple cameras is *Maximum a Posteriori* Tracking [65] (defined in [14]). A few other RE-ID works have exploited the temporal statistics of people moving from camera to camera [54, 88].

2.3 EVALUATION OF RE-ID ALGORITHMS

2.3.1 *Datasets*

During the development and evaluation of RE-ID algorithms it is essential to rely on a properly annotated dataset (a survey of the most relevant datasets for RE-ID is done below). Datasets typically include the location and identity of persons on the camera network images, annotated by humans, that can be used as "ground truth" to evaluate the accuracy of the developed algorithms. The minimum required is a dataset composed of cropped images of people annotated with their respective identifications. If the dataset also makes available from which camera each image is provenient, it is possible to guarantee that the gallery images come solely from some cameras and the probe images solely from other cameras. The availability of synchronized video data instead of unordered frames allows the exploitation of temporal constraints to reduce complexity in the classification stage [88]. As further developed in this thesis, temporal constraints significantly improve the performance of RE-ID algorithms. Also, an unique dataset was developed in-house that provides synchronized video data [116]).

Datasets most commonly offer rectangular images of pedestrians cropped from a bounding box. This raises the issue of unwanted background. Each bounding box will contain some amount of background that it is assumed to be uncorrelated with each persons' appearance, and therefore unwanted. If the dataset provides foreground masks, this issue is sidestepped. When the video data is provided in the dataset, some kind of background subtraction may be used prior to feature extraction. Otherwise, algorithms such as body-part detection may still be applied and features may be extracted only from these local regions in the pedestrian image. What was used in the course of this work is detailed below in Section 2.2.1.

Here it's described several datasets that have been developed over the years, and that will be used in this thesis experiments. Table 3 offers a comparison of the datasets' relevant parameters, and they are described in more detail below:

ETHZ4REID [1] This dataset presented in [112] was created from the more general ETHZ dataset [40][2]. It is composed of cropped images from 3 video sequences, captured by a single head-height moving camera. With more than 7500 images of about 150 pedestrians, it provides the notable challenge of occlusions and illumination changes, while having very little pose variation.

Figure 4: Sample images from the ETHZ dataset. It contains a lot of images of each pedestrian from a single camera view at head level, in a city street.

VIPER [3] It presents the challenges of different poses, viewpoints and lighting conditions. This dataset was presented in [58] in 2008. It remained one of the most challenging single-shot dataset up until 2013 when the HDA dataset [116] was released. It contains only 2 images of 632 people, each in a different pose.

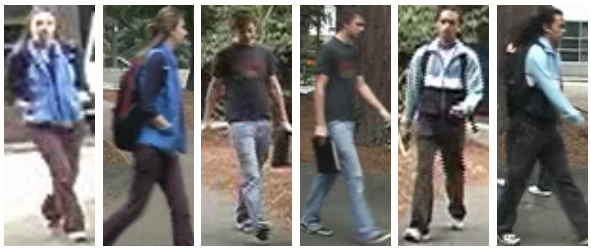


Figure 5: Sample images from the VIPeR dataset. It has only two images for each pedestrian, from two distinct cameras, in an outdoors environment. Almost all pairs have the respective pedestrian in different poses, facing different directions with about a 90° different angle.

ILIDS4REID This dataset is composed by 476 images of 119 people. It was presented in [127] and built from the iLIDS Multiple-Camera Tracking

1 ETHZ4REID downloadable at http://homepages.dcc.ufmg.br/~william/datasets.html
2 ETHZ downloadable at https://data.vision.ee.ethz.ch/cvl/aess/dataset/
3 VIPeR downloadable at http://vision.soe.ucsc.edu/node/178

Scenario, which was captured in a busy airport hall. The notable challenges it presents are the presence of occlusions and large illumination changes.

Figure 6: Sample images from the iLIDS4REID dataset. It contains a few images of each pedestrian from up to two different camera views in an airport.

CAVIAR4REID This dataset was made with images extracted from the more general CAVIAR dataset[4]. It was presented in [25], and it includes the views of two cameras in a shopping center, with overlapping field of view at 90° angle. It contains 10 images per camera per individual of 50 persons and 10 more images per person of 22 pedestrians that only appear in one of the cameras.

Figure 7: Sample images from the CAVIAR4REID dataset. It contains a ten images of each pedestrian from each camera, from up to two cameras in a shopping center with very low resolution.

3DPES [5] This dataset was presented in [13], and was the first re-identification dataset with more than 2 cameras. It contains images from 8 fixed cameras with non-overlapping fields of view, of 200 different people, with a small and variable number of detections per person (1000 in total).

PRID2011 [6] Created in co-operation with the Austrian Institute of Technology, presented in [63], it gives us two camera views from above of pedestrians walking in the street. 200 pedestrians were captured by both cameras, and over 700 other people were only captured by one or the other

4 CAVIAR downloadable at http://homepages.inf.ed.ac.uk/rbf/CAVIAR/
5 3DPeS downloadable at http://www.openvisor.org/3dpes.asp
6 PRID2011 downloadable at http://lrs.icg.tugraz.at/datasets/prid/

Figure 8: Sample images from the 3DPeS dataset. It contains some images from up to eight different camera views in a college campus environment.

camera. All pedestrians have at least 5 cropped images and some have many more (up to 80).

Figure 9: Sample images of a single person from the PRID2011 dataset. It contains many images of each pedestrian, from two camera views looking at a street crossing.

ILIDS-MA [7] Two datasets were presented in [12], this one and iLIDS-AA. This dataset contains 40 people, each with 42 manually annotated images, was also recorded in an airport hall.

ILIDS-AA [7] This companion dataset, also presented in [12], is similar to iLIDS-MA. It has a variable number of images for each of its 100 people (10500 in total). These images are not manually annotated, but instead cropped with a background subtraction algorithm, yielding the notable challenge of pedestrian detections with non-centered bounding boxes.

ILIDS-VID [8] This dataset was recently presented in [120] and was also built from the iLIDS Multiple-Camera Tracking Scenario. It contains 43'800 cropped images of 300 people visible from two cameras.

7 iLIDS-MA and AA downloadable at http://www-sop.inria.fr/members/Slawomir.Bak/gpEasy/DataSet

8 iLIDS-VID downloadable at http://www.eecs.qmul.ac.uk/~xz303/downloads_qmul_iLIDS-VID_ReID_dataset.html

Figure 10: Sample images of a single person from the iLIDS-MA dataset. It contains many images of each pedestrian, from two camera views.

Figure 11: Sample images of a single person from the iLIDS-AA dataset. It contains many images of each pedestrian, from two camera views in an airport, with the notable challenge of automatically generated unreliable bounding boxes.

HDA [9] The most notable re-identification dataset, developed in Vislab-Lisbon [116]. It is a dataset of 18 cameras with almost no overlapping fields of view. It contains a large and variable number of detections per each of its 85 persons (over 64'000 in total). With the notable characteristic of including high-definition images from 11 of the 20 cameras – one 4 mega pixel camera (2560×1600 resolution), and ten 1 mega pixel cameras (1280×800 resolution) – and one over-head camera. The presence of harsh illumination changes, very large scale changes (due to the HD cameras), severe occlusions, the fact that several subjects change clothes from one view to the next (*i.e.,* put on jackets), and the presence of one over-head camera make it one of the most challenging re-identification datasets up to date. The label set is very complete and includes pedestrians with large occlusions that no algorithm to date is able to detect, but this provide a very complete and challenging benchmark set for the current and yet to come person detection and re-identification systems.

9 You can request to download HDA at vislab.isr.ist.utl.pt/hda-dataset/

Figure 12: Sample images of a single person from the iLIDS-VID dataset. It contains many images of each pedestrian, from two camera views in an airport.

2.3.2 *Evaluation Metrics*

The standard metric for Re-Identification (RE-ID) evaluation is the Cumulative Matching Characteristic curve (CMC), that shows how often, on average, the correct person ID is included in the best r matches against the gallery set for each probe image. However, since the CMC computes the average re-identification rate for the probes evaluated, it ignores by design the Missed Detections (MDs) introduced by the Pedestrian Detection (PD) algorithm. This implies that other metrics should be used to complement the CMC when evaluating and integrated detection and classification system.

In other fields such as object detection and tracking, precision and recall metrics are used to evaluate the algorithms[10]. Recall encodes how many relevant samples were recovered by the system. Precision encodes how many of the recovered samples were relevant. In this work these metrics are adapted to evaluate the integrated detection and classification system (see Section 4.4.1).

10 Such as in the iLIDS dataset's user guide: http://www.siaonline.org/SiteAssets/Standards/PerimeterSecurity/iLidsUserGuide.pdf

Figure 13: Sample images from the HDA dataset. It contains many images of very different scales, from VGA up to 4MPixel cameras. From up to thirteen different camera views in a office space environment. It includes the notable challenge of changing apparel.

Name	#CA	#SE	#FR	#BB	#PE	Max. Res.	Main Applications	**DC**	**SV**
ETHZ4REID [40]	1	0	0	8580	146	453×226 (C)	RE-ID	-	-
VIPeR [58]	2	0	0	1264	632	128×48 (C)	RE-ID	✓	×
iLIDS4REID [127]	2	0	0	476	119	304×153 (C)	RE-ID	×	×
CAVIAR4REID [25]	2	0	0	1220	72	384×288	RE-ID	✓	×
3DPeS [13]	8	0	0	605	199	280×141 (C)	RE-ID	×	×
PRID2011 [63]	2	0	0	94988	934	128×64 (C)	RE-ID	✓	×
iLIDS-MA [12]	2	0	0	3680	40	589×294 (C)	RE-ID	✓	×
iLIDS-AA [12]	2	0	0	10329	100	118×238 (C)	RE-ID	✓	×
iLIDS-VID [120]	2	0	0	43800	300	128×64 (C)	RE-ID	✓	×
HDA [116]	13	13	75207	64028	85	2560×1600	PD, RE-ID, Tracking	✓	✓

Table 3: Main characteristics of the surveyed data sets. I compare the number of cameras in the dataset (#CA), the number of video sequences (#SE), the number of video frames (#FR), the number of person bounding box labels (#BB), the number of person identity labels (#PE), the maximum video resolution available, and the main application envisaged for the data set. Data sets whose number of video sequences is 0 are composed of independent photographies. Data sets providing cropped images instead of full frames are indicated with 0 in the number of frames. In these cases the maximum resolution refers to the size of the cropped images and is followed by symbol (C). None provide foreground pixel masks. Finally it is noted if the dataset gives information of from which camera each image is provenient under **DC** and if it gives synchronization information between cameras (**SV**).

3

RE-IDENTIFICATION IN CAMERA NETWORKS

This chapter presents the solutions developed during the thesis towards the improvement of re-identification systems. First we look at the integration of automatic pedestrian detectors [114, 55]) that hamper RE-ID, when both are integrated. with RE-ID algorithms. The following goal is to describe the advances in descriptor extraction and features, proposed in this work. Afterwards, describe the advanced Multi-View classification algorithm. Finally, expound the novel metrics proposed to properly access real-world RE-ID systems.

3.1 INTEGRATION WITH PEDESTRIAN DETECTOR

For almost all Re-Identification (RE-ID) algorithms in the state of the art, the data for the RE-ID problem is provided in the shape of hand-cropped Bounding Box (BB), rather than in the shape of full image frames. Such BBs are centered around fully visible, upright persons, and the focus of the RE-ID algorithms is on the feature extraction and BB classification. This means standard RE-ID state of the art works assume perfect pedestrian detection.

However, the purpose of an automated RE-ID system is that of re-identifying people directly in images, without requiring manual intervention to produce the BBs. [12] is one of the few works to have performed re-identification with not-manually-cropped person images. Here the authors use a background subtraction method to create the bounding boxes, but then manually pick which BBs to use, only addressing the issue of unreliable bounding boxes, and not the false positives and missed detections. Methods that actively integrate pedestrian detection and re-identification, or works that propose metrics to evaluate integrated RE-ID systems are even scarcer in the literature.

The works that relate the most with this part of this thesis are [95] and [67]. In [95], the system's full flow (*i.e.*, pedestrian detection and re-identification) is presented with a transient gallery to tackle open scenarios. They use RGB-D data, that with the current technology has a range limit of 5 meters, which may be limiting in some environments, and only employ one camera, attempting to recognize the same pedestrian in several passes in front of the camera. In [67] an approach that integrates PD and RE-ID is presented, using infrared images from the CASIA Gait database [33].

However, in those works, the performance is evaluated on the overall system, not being possible to ascertain the impact of integrating each constituent part in the system. Furthermore, important issues such as how re-identification performance is penalized when pedestrian detection or tracking failures exist are not evaluated. One goal of this work is precisely to investigate how to enhance the link between pedestrian detection and re-identification algorithms to improve the overall performance.

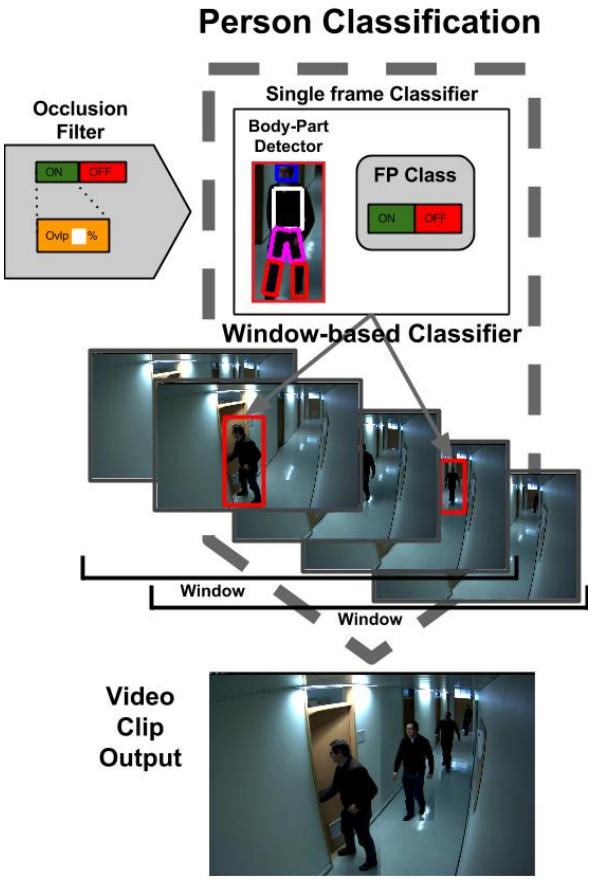


Figure 14: The Person Classification block seen in Figure 2 is here expanded to highlight the novelties to the proposed re-identification system architecture. First, the bounding boxes provided by the pedestrian detection stage are optionally processed by the Occlusion Filter (gray block on the left) that discards occluded samples (samples under occlusion with over a threshold of occlusion). Then, body-part detection is ran in the provided bounding boxes (as can be seen just right of the Occlusion Filter) so features are extracted from those local regions. The Single frame Classifier block represents any classification algorithm that takes features and classify them into people classes. Here, the second gray block represents an additional class: a class to model the false positive samples, can be used by the classifier. This deals with the spurious false positive detections that inevitably appear with an automatic person detector. The window-based classifier (the dashed line that encompasses the single frame classifier) then takes the classifications, and if there are enough positive re-identifications in one or more temporal windows, outputs a video-clip with the combination of such windows.

Integrating PD and RE-ID poses several challenges. Detecting people in images is a hard task, in fact even the best detectors in the state of the art are subject to the production of at least two types of errors: False Positive (FP) and Missed Detection (MD). Such errors have a direct impact on the performance of the compounded system, that is, FPs generate BBs which are impossible for the system to correctly classify as one of the persons in the galley set. MDs, on the other hand, cause an individual to simply go undetected, and thus, unclassified. Even the correctly detected persons may give rise to the following difficulties: (1) the PD algorithm can generate a BB not centered around the person or at a non-optimal scale – this might hinder the feature extraction phase, prior to the classification, (2) the detected person may be partially occluded, yet again hampering feature extraction, and finally (3) there can be the case of detecting people who are not part of the RE-ID gallery set, posing an issue similar to that of FPs, *i.e.*, there is no correct class that the system can assign them.

This work focuses on the closed-space scenario (explained in Section 1.3) while tackling the above mentioned difficulties. In the remainder of this section several additional modules to the proposed RE-ID architecture (Figure 2) are described. These modules, illustrated in Figure 14 solve some of the aforementioned issues, such as (i) body-part detection to ameliorate the issue of unreliable bounding boxes, (ii) a window-based classifier to filter the single-frame classifier output thus reducing classification errors, (iii) collating the output frames into short video-clips which reduces the operator's attentional load and also re-captures some missed detections, (iv) an occlusion filter to deal with occluded detections, and (v) a false positive class to deal with non-people detections.

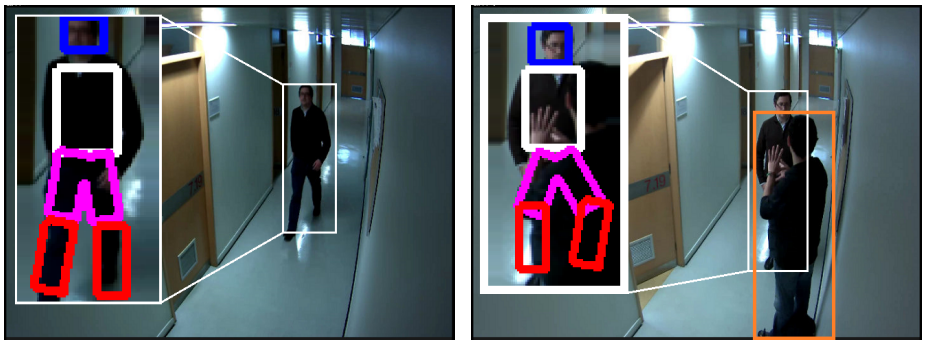
3.1.1 *Body-Part Detection for Feature Extraction Alignment*

The issue of bounding box misalignment can be ameliorated by applying PS [4] to detect body parts inside the BBs. Features will then be extracted from the relevant image regions (the body-parts) even if the BB is not correctly centered or scaled to the pedestrian (see Section 3.2 for more detail).

3.1.2 *Occlusion Filter*

This module was created in collaboration with colleagues [115], it is described here for completeness.

As mentioned above, the RE-ID performance can be jeopardized by incorporating detections of occluded pedestrians. Therefore, the Occlusion Filter was devised. It is a filtering block between the PD and the RE-ID modules (see Figure 14), with the intent of improving the RE-ID performance. The Occlusion Filter uses geometrical reasoning to reject BBs which can harm the performance of the RE-ID stage (BBs depicting partially occluded people). A Bounding Box (BB) including a person appearing under partial occlusion generates features different from a BB including the same person under full visibility



(a) (b)

Figure 15: Example of body-part detection for feature extraction in two instances: (a) a person appearing with full visibility and (b) under partial occlusion, with detected bounding boxes overlap. The feature extraction on the occluded person mistakenly extracts some features from the occluding pedestrian.

conditions. When the partial occlusion is caused by a second person standing between the camera and the original person, the extracted features can be a mixture of those generated by the two people, making the identity classification especially hard (see illustration in Figure 15). For this reason, it would be advantageous for the RE-ID module to receive only BBs depicting fully visible people.

Though the visibility information is not available to the system, it can be estimated quite accurately with a scene geometry reasoning: in a typical scenario the camera's perspective projection makes proximal pedestrians extend to relatively lower regions of the image. Thus, the filter computes the overlap among all pairs of detections in one image and rejects the one in each overlapping pair for which the lower side of the BB is higher (as illustrated in Figure 16). Considering the mismatch between the shape of the pedestrians' bodies and that of the BBs, it is clear that an overlap between BBs does not always imply an overlap between the corresponding pedestrians' projections on the image. An overlap threshold for the filter is defined, considering as overlapping only detections whose overlap is above such threshold. The impact of the overlap threshold on the RE-ID performance was analyzed in [115], where the optimal value of 30% was proposed.

3.1.3 *False Positives Class*

Another contribution of this work is to adapt the classification stage so that it can deal with the FPs produced by the PD. The standard RE-ID module cannot deal properly with FPs: each FP turns into a wrongly classified instance for the RE-ID. Observing that the appearance of the FPs in a given scenario is not completely random, but is worth modeling (see Figure 17), a FP class is introduced for the RE-ID module. In these conditions, a correct output exists for when a FP is presented on the RE-ID's input: the FP class. This change allows us to coherently evaluate the performance of the integrated system.

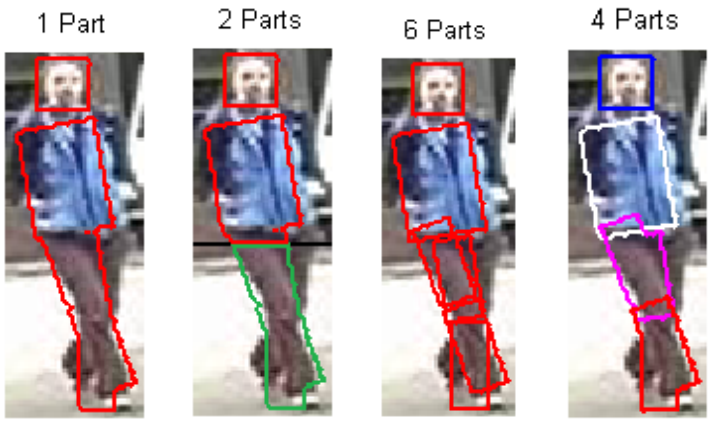
Figure 16: An example of geometrical reasoning: two detection bounding boxes overlap. The comparison between the lower sides of the two bounding boxes leads to the conclusion that the person marked with the red, dashed bounding box is occluded by the person in the green, continuous bounding box. Therefore the corresponding bounding box is rejected.

Figure 17: Example False Positive samples in the False Positive Class training set.

3.2 BODY-PART DETECTION FOR DESCRIPTOR EXTRACTION

In the beginning of this work a background subtraction algorithm (LOTS [21]) was used to detect pedestrians, detect the foreground pixels, and thus extract features from the pixels of each full body detection. A new method was then proposed, extracting features from two body areas, separated by the waist of the pedestrian (as illustrated in Figure 18b) that improved the performance of re-identification [44] (results in Section 4.1). Later on Cheng *et al.* [25] verified that applying PS [4] to further detect body parts (head, torso, 2 thighs, 2 fore-legs) and extract features from those 6 separate areas further improves re-identification results (results in Section 4.1).

My final contribution to the descriptor extraction part of the RE-ID problem consists of realizing that one thigh area is not usually distinguishable from the other thigh area, nor is one fore-leg area normally different from the other, and thus these twin areas should not be considered separately from one another, and definitely should not be ordered (as was done in [25]). Thus, I verified that not dividing each leg and fore-leg into separate and ordered regions in the feature vector slightly increases RE-ID results (see Section 4.1 in the next chapter).



(a) Full-body detection (1 Part).

(b) Example of waist detection and division (2 Parts).

(c) Head, torso, 2 thighs and 2 fore-legs detection (6 Parts).

(d) Head, torso, thighs and fore-legs detection (4 Parts).

Figure 18: Visual examples of the different ways one can extract descriptors from a pedestrian detection. (18a) shows the baseline – full body feature extraction; (18b) depicts detecting the waist of a pedestrian and extracting features from the upper body and lower body separately [44]; (18c) illustrates Cheng *et al.* [25] work, which applies PS to detect body parts; (18d) applies PS to detect body parts, then joins the detections of the separate thighs in one region, the detection of the separate fore-legs in another region (this work). In the results chapter it will be shown that from (a) to (d) the performance increases monotonically.

This also ameliorates the issue of bounding boxes not exactly centered or not exactly scaled to the pedestrian size, as mentioned above (see Section 3.1.1).

3.3 CLASSIFICATION

This section describes the contributions put forth in the area of classification for re-identification. First a semi-supervised multi-feature integration classification algorithm was adapted for RE-ID. Then a wrapper for single-frame classification was developed to filter out spurious mis-classification and recapture some missed detections (see Section 3.3.2). Finally, some attention was given to the concern of the operating user attentional load, by combining single-frame output into small video-clips (see Section 3.3.3).

3.3.1 *Multi-View Classification*

This section is devoted to the presentation and discussion of Multi-View classification, a mathematical formulation to train a classifier integrating several features [46]. Many features are only useful in parts of the data, *e.g.*, texture

features are mostly useful in pedestrian with textured clothes, and color features are usually less useful in the pedestrian legs where the pants tend to be of similar colors. Many features should be used to increase the ability to discriminate between the numerous pedestrians, some very similar. This necessitates a good feature integration method, which is the focus of this section.

Multi-View is a semi-supervised algorithm. It is built to exploit both the information available in labeled and unlabeled samples, which is useful when there are very few labeled samples as is common in the RE-ID problem. In the absence of unlabeled data it acts as a supervised algorithm. In the supervised object recognition field, multi-view is compared favorably with other works [124, 27, 119] as shown in [105]. It has been successfully applied to the Object Recognition and Bird Categorization problems [105], as well as to my previous work on the Re-Identification problem [46] (where the test samples were used as unlabeled data).

While a regular classifier takes a feature vector and outputs one label for it, Multi-View splits the data in several views, trains several classifiers from that same data, and then fuses them. The core of Multi-View is the exploitation of complementary information available on each view. The Multi-View formulation assumes that each view is "sufficient" to train a "good" classifier (above chance). It also assumes that the feature split in several views actually exists and that the data in each view is *conditionally independent*. This implies that the classifiers will perform differently on different parts of the data. These classifiers (one per view) will be trained together, and joined to produce a final better classifier. The data is separated in labeled and unlabeled samples. Multi-View, during training, teaches all classifiers to correctly classify the labeled samples, and promotes concordance between the classifiers in all the samples (labeled and unlabeled). Since the classifiers are assumed to be at least "good", this concordance is expected to agree more often on correct classifications that on incorrect ones, pushing all classifiers to be better than they would be if trained alone.

In Multi-View, each view represents different facets of a given sample. They can be different features (*e.g.*, color, texture), or attributes (*e.g.*, hair length, gender, income level) of the sample, or they can be different inputs of the same sample (*i.e.*, several images taken from a same camera, or several images taken from different cameras). They can be either vectors, matrices or sets; of real numbers, integers, or other symbols[1]. In the results chapter of this work, in most experiments, views represent feature vectors such as those in Figure 22, extracted from the person descriptors described in Section 3.2. In

1 Lodhi *et al.* [87] describes how to construct kernels to compare strings. Kernel functions can be defined over general sets, by assigning to each pair of elements (strings, graphs, images) an 'inner product' in a feature space. If 'inner product' is clearly defined for the symbols in use, any of the general purpose kernels can be used. One successfully used feature on strings is the frequency of words (usually after removing stop-words and the inflection of words). *E.g.*, in [87] Lodhi's feature space is the set of all (non-contiguous) substrings of k-symbols. The more substrings two documents have in common, the more similar they are and the higher their inner product.

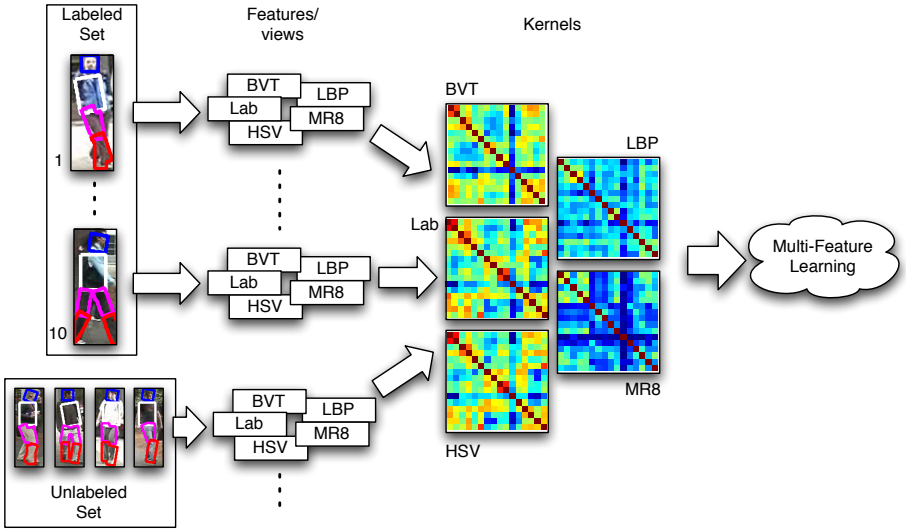


Figure 19: Overview of the proposed classification method. The features of each individual are extracted from labeled and unlabeled images. Then, kernels are computed for each feature and finally the multi-view classifier is trained.

one experiment of a Multi-Shot scenario, each view will represent one image. In the rest of the text we'll use the terms features and views interchangeably.

The pipeline of the proposed method is depicted in Figure 19. The feature descriptors of each individual in the *labeled* and *unlabeled* sets are extracted from detected body parts. Then, the similarity between the descriptors is computed for each feature by mean of kernel operators (more detail below). Multi-view learning consists in estimating the parameters of the classifiers given the training set (see the following sections). Given a probe image, the testing phase consists in computing the similarity of each descriptor with the gallery samples and use the learned parameters to classify it.

3.3.1.1 *Multi-View Learning*

In this work, the multi-view learning framework [105] is applied to the re-identification problem with the *views* corresponding to different types of *fea-*

Notation:
I_p: identity matrix of size $p \times p$
$\otimes$: Kronecker product
$\mathbf{e}_m$: vector of ones of size m
C^*: Complex conjugate of C
Lower case letters for single numbers (*e.g.*, number of views m).
Bold lower case letters for vectors (*e.g.*, feature vector $\mathbf{x}_i^j$ of view j of sample i).
Capital letters for matrices (*e.g.*, combination operator C).
Bold capital letters for matrices of matrices (block matrices).
Caligraphic capital case letters for sets (*e.g.*, feature set of the ith sample $\mathcal{X}_i$).
Bold caligraphic capital case letters for sets of sets.

ture vectors[2] (color and textures) extracted from the images as indicated in Section 3.2.

Suppose there is a training set $\{(\mathcal{X}_i, \mathbf{y}_i)\}_{i=1}^{l} \cup \{\mathcal{X}_i\}_{i=l+1}^{l+u}$, where $\mathcal{X}_i$ represents the set of m views represented by features vectors $\mathbf{x}_i^j$, $j = 1 \dots m$, extracted from the i-th image in the training set. These feature vectors are of size $\mathbb{R}^{d_j}$, where d_j is the dimension of view j. To each sample $\mathcal{X}_i$ corresponds an identity label $\mathbf{y}_i$. The set on the left of the union symbol is called the labeled set with l samples, while the one on the right is called the unlabeled set with u samples, in which the ground truth labels $\mathbf{y}_i$ are not available. In re-identification, the labeled set corresponds the gallery, and the unlabeled set contains data acquired in the same conditions but without labels. If the unlabeled set is not available, the method performs supervised learning. The unlabeled data set has the purpose of providing more structure to the data, which helps on the learning process.

Given that p is the number of identities in the re-identification problem, each identity label $\mathbf{y}_i$, $1 \leqslant i \leqslant l$, has the form $\mathbf{y}_i = [-1 \cdots 1 \cdots -1]^T$. It is a vector of -1's with a single 1 at the p-th location if $\mathcal{X}_i$ is in the p-th class. Finally, the output of each classifier is a column vector in $\mathbb{R}^p$.

Under a multi-view formulation, each of the m classifiers learned (one per view) will have the following form:

$$\mathbf{f}^m(\mathbf{x}_j^m) = \sum_{i=1}^{l+u} k^m(\mathbf{x}_j^m, \mathbf{x}_i^m)\mathbf{a}_i^m \in \mathbb{R}^p \tag{1}$$

Here k^m is a kernel that induces a Reproducing Kernel Hilbert Space (RKHS)[3] $\mathcal{H}_K$ of functions $\mathbf{f}^m$, that receives two feature vectors of view m and outputs a scalar (valid kernels listed in Section 3.3.1.3). $\mathbf{a}_i^m$ are vectors in $\mathbb{R}^p$ of weights to be learned. These vectors $\mathbf{a}_i^m$ will weight each view of each training sample.

The view classifier outputs are then linearly combined. If the view classifier outputs are concatenated in a long vector

$$\mathbf{f}(\mathcal{X}_i) = [\mathbf{f}^1(\mathbf{x}_i^1)^T, \dots, \mathbf{f}^m(\mathbf{x}_i^m)^T]^T \in \mathbb{R}^{p \cdot m},$$

the linear combination can be represented in matrix form:

$$C = 1/m \cdot [I_p \dots I_p] \in \mathbb{R}^{p \times p \cdot m}$$
$$C\mathbf{f}(\mathcal{X}) = \frac{1}{m}(\mathbf{f}^1(\mathbf{x}^1) + \cdots + \mathbf{f}^m(\mathbf{x}^m)) \in \mathbb{R}^p.$$

where C is the concatenation of m p-sized diagonal matrices with $1/m$ in the diagonal.

2 Multi-view is an algorithm that can take any set of aspects of a sample as views, and in the next chapter features extracted from a single image are used as views for many experiments. For one multi-shot experiment different images of a same pedestrian are used as the source of each view (Section 4.2.5).

3 see [7] for the definition and details on Reproducing Kernel Hilbert Spaces (RKHSs).

Given the training set, re-identification under a multi-view formulation consists of the following optimization problem based on the least square loss function:

$$\min_{\mathbf{f}\in\mathcal{H}_K} \frac{1}{l}\sum_{i=1}^{l}\left\|\mathbf{y}_i - C\mathbf{f}(\mathcal{X}_i)\right\|^2 + \gamma_A\sum_{i=1}^{l+u}\|\mathbf{f}(\mathcal{X}_i)\|^2 + \gamma_I\sum_{i=1}^{l+u}\sum_{j,k=1,j<k}^{m}\left\|\mathbf{f}^j(\mathbf{x}_i^j) - \mathbf{f}^k(\mathbf{x}_i^k)\right\|^2 \tag{2}$$

where the regularization parameter γ_A must be strictly positive and $\gamma_I \geqslant 0$.

The first term of Equation 2 is the least square loss function that measures the error between the final output $C\mathbf{f}(\mathcal{X}_i)$ for $\mathcal{X}_i$ with the given label $\mathbf{y}_i$, for each i. The main difference with the standard least square optimization is that this formulation combines the different views. In particular, if each input instance $\mathcal{X}$ has many views, then $\mathbf{f}(\mathcal{X})$ represents the output values from all the views. These values are combined by the operator C to give the final output value.

The second summand is the standard RKHS regularization term. It exists to minimize the classifiers parameters and therefore to improve its generalization power. Intuitively, when there are "rare" samples, these samples correlate very highly with some features that don't necessarily have high predictive power in general. If this generalization term was not present, those correlations would cause the output to increase dramatically at those "rare" samples leading to worse performance outside the training data. This is the effect of overfitting.

The third summand is the multi-feature manifold regularization [105], which performs consistency regularization across different views. It penalizes non-consensus between the different classifiers. This is what promotes the concordance between the classifiers in as much samples as possible. This is also the reason Multi-View requires the assumption that each view is "sufficient" to train a "good" classifier, so that most classifiers more often classify correctly and thus push the remainder classifiers to better performance levels.

3.3.1.2 *Solution to the optimization problem*

Problem (2) is an instance of unconstrained quadratic optimization on the classifier coefficients a_i^m. It can be solved by finding the stationarity points of (2). This can be achieved by solving for the points where the derivatives of (2) equate to zero. This was done in [105], and I will re-do the derivation here for completeness. Lets first rewrite (2) in matrix form to simplify the algebraic derivation.

$$\sum_{i=1}^{l+u}\sum_{j,k=1,j<k}^{m}\left\|\mathbf{f}^j(\mathbf{x}_i^j) - \mathbf{f}^k(\mathbf{x}_i^k)\right\|^2 = \sum_{i=1}^{l+u}\mathbf{f}(\mathcal{X}_i)^T M\mathbf{f}(\mathcal{X}_i),$$

for

$$M = \begin{bmatrix} p\times m-1 & -1 & \cdots & -1 \\ -1 & p\times m-1 & \cdots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \cdots & p\times m-1 \end{bmatrix} \in \mathbb{R}^{p\cdot m\times p\cdot m}$$

and if all the $\mathbf{f}(X_i)$ are concatenated into a long vector $\boldsymbol{f}$

$$\boldsymbol{f} = \left[\mathbf{f}(X_1)^T,\ldots,\mathbf{f}(X_l)^T,\ldots,\mathbf{f}(X_{l+u})^T\right]^T \qquad \in\mathbb{R}^{p\cdot m\cdot(l+u)} \tag{3}$$

then

$$\sum_{i=1}^{l+u}\mathbf{f}(X_i)^T M\mathbf{f}(X_i) = \boldsymbol{f}^T\mathbf{M}\boldsymbol{f}$$

for $\mathbf{M}$ being a block matrix with blocks M in its diagonal:

$$\mathbf{M} = \begin{bmatrix} [M] & [0] & \cdots & [0] \\ [0] & [M] & \cdots & [0] \\ \vdots & \vdots & \ddots & \vdots \\ [0] & [0] & \cdots & [M] \end{bmatrix} \in \mathbb{R}^{(l+u)\cdot p\cdot m\times(l+u)\cdot p\cdot m}$$

The same $\boldsymbol{f}$ defined in Equation 3 can be used to simplify the second term thusly:

$$\sum_{i=1}^{l+u}\|\mathbf{f}(X_i)\|^2 = \|\boldsymbol{f}\|^2$$

Finally, by concatenating all $\mathbf{y}_i$ into one long $\mathbf{y}$ vector, with zeros in the unlabeled samples respective positions

$$\mathbf{y} = \left[\mathbf{y}_1^T,\ldots,\mathbf{y}_l^T,0,\ldots,0\right]^T \qquad \in\mathbb{R}^{p\cdot(l+u)}$$

and doing similarly for C

$$\mathbf{C} = [[C],\ldots,[C],[0],\ldots,[0]] \qquad \in\mathbb{R}^{p\times p\cdot m\cdot(l+u)}.$$

it is possible to write Equation 2 as

$$\min_{f\in\mathcal{H}_K}\frac{1}{l}\|\mathbf{y}-\mathbf{C}\boldsymbol{f}\|^2+\gamma_A\|\boldsymbol{f}\|^2+\gamma_I\boldsymbol{f}^T\mathbf{M}\boldsymbol{f}$$

Expanding the norms yields

$$\min_{f\in\mathcal{H}_K}\frac{1}{l}\left(\mathbf{y}^T\mathbf{y}-\mathbf{y}^T\mathbf{C}\boldsymbol{f}-\boldsymbol{f}^T\mathbf{C}^T\mathbf{y}+\boldsymbol{f}^T\mathbf{C}^T\mathbf{C}\boldsymbol{f}\right)+\gamma_A\boldsymbol{f}^T\boldsymbol{f}+\gamma_I\boldsymbol{f}^T\mathbf{M}\boldsymbol{f}.$$

Since $\mathbf{y}^T\mathbf{C}\boldsymbol{f}$ and $\boldsymbol{f}^T\mathbf{C}^T\mathbf{y}$ are scalar, they are equal to their transposed and thus

$$\boldsymbol{f}^T\mathbf{C}^T\mathbf{y} = \left(\boldsymbol{f}^T\mathbf{C}^T\mathbf{y}\right)^T = \mathbf{y}^T\mathbf{C}\boldsymbol{f}$$

So, Equation 2 becomes:

$$\min_{f\in\mathcal{H}_K} \frac{1}{l}\left(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{C}\boldsymbol{f} + \boldsymbol{f}^T\mathbf{C}^T\mathbf{C}\boldsymbol{f}\right) + \gamma_A\boldsymbol{f}^T\boldsymbol{f} + \gamma_I\boldsymbol{f}^T\mathbf{M}\boldsymbol{f} \tag{4}$$

Differentiating in order to $\boldsymbol{f}$ yields

$$\begin{aligned}
&\frac{\partial(\mathbf{y}^T\mathbf{y})}{\partial\boldsymbol{f}} = 0\\
&\frac{\partial(-2\mathbf{y}^T\mathbf{C}\boldsymbol{f})}{\partial\boldsymbol{f}} = -2\mathbf{y}^T\mathbf{C}\\
&\frac{\partial(\boldsymbol{f}^T\mathbf{C}^T\mathbf{C}\boldsymbol{f})}{\partial\boldsymbol{f}} = \boldsymbol{f}^T(\mathbf{C}^T\mathbf{C} + (\mathbf{C}^T\mathbf{C})^T) = \boldsymbol{f}^T(\mathbf{C}^T\mathbf{C} + \mathbf{C}^T\mathbf{C}) = 2\boldsymbol{f}^T\mathbf{C}^T\mathbf{C}\\
&\frac{\partial(\gamma_A\boldsymbol{f}^T\boldsymbol{f})}{\partial\boldsymbol{f}} = 2\gamma_A\boldsymbol{f}^T\\
&\frac{\partial(\gamma_I\boldsymbol{f}^T\mathbf{M}\boldsymbol{f})}{\partial\boldsymbol{f}} = \gamma_I\boldsymbol{f}^T(\mathbf{M} + \mathbf{M}^T)
\end{aligned}$$

Because $\mathbf{M}$ is symmetric $\mathbf{M}^T = \mathbf{M}$, thus $\gamma_I\boldsymbol{f}^T(\mathbf{M} + \mathbf{M}^T) = 2\gamma_I\boldsymbol{f}^T\mathbf{M}$. Thus differentiating Equation 4 and equating to zero yields

$$\frac{1}{l}(-2\mathbf{y}^T\mathbf{C} + 2\boldsymbol{f}^T\mathbf{C}^T\mathbf{C}) + 2\gamma_A\boldsymbol{f}^T + 2\gamma_I\boldsymbol{f}^T\mathbf{M} = 0$$

which is equivalent to

$$\begin{aligned}
&\frac{1}{l}\boldsymbol{f}^T\mathbf{C}^T\mathbf{C} + \gamma_A\boldsymbol{f}^T + \gamma_I\boldsymbol{f}^T\mathbf{M} = \frac{1}{l}\mathbf{y}^T\mathbf{C}\\
&\boldsymbol{f}^T\left(\frac{1}{l}\mathbf{C}^T\mathbf{C} + \gamma_A I + \gamma_I\mathbf{M}\right) = \frac{1}{l}\mathbf{y}^T\mathbf{C}
\end{aligned}$$

$$\left(\frac{1}{l}\mathbf{C}^T\mathbf{C} + \gamma_A I + \gamma_I\mathbf{M}\right)^T\boldsymbol{f} = \frac{1}{l}\mathbf{C}^T\mathbf{y} \tag{5}$$

Now, given Equation 1, each view classifier $\mathbf{f}^m$ can be written in matrix form as follows:

$$\mathbf{f}^m(.) = \sum_{i=1}^{l+u} k^m(., \mathbf{x}_i^m)\,\mathbf{a}_i^m = K^m(.)\mathbf{a}^m \qquad \in\mathbb{R}^p$$

where

$$K^m(.) = \begin{bmatrix} k^m(., \mathbf{x}_1^m)\cdot I_p & \dots & k^m(., \mathbf{x}_{l+u}^m)\cdot I_p \end{bmatrix} \qquad \in\mathbb{R}^{p\times p\cdot(l+u)}$$

and $\mathbf{a}^m$ is the concatenation of all the $\mathbf{a}_i^m$ into a column vector

$$\mathbf{a}^m = \left[\mathbf{a}_1^m \dots \mathbf{a}_{l+u}^m\right]^T \qquad \in \mathbb{R}^{p \cdot (l+u)}$$

Then, the concatenation of each view classifier $\mathbf{f}(\mathcal{X}_i) = \left[\mathbf{f}^1(\mathbf{x}_i^1)^T, \dots, \mathbf{f}^m(\mathbf{x}_i^m)^T\right]^T$ can be written in matrix form as follows

$$\mathbf{f}(\mathcal{X}_i) = [\mathbf{f}^1(\mathbf{x}_i^1), \dots, \mathbf{f}^m(\mathbf{x}_i^m)]^T = \mathbf{K}(\mathcal{X}_i)\mathbf{a} \qquad \in \mathbb{R}^{m \cdot p}$$

where $\mathbf{K}(\mathcal{X}_i)$ is a block matrix with

$$\mathbf{K}(\mathcal{X}_i) = \begin{bmatrix} \left[K^1(\mathbf{x}_i^1)\right] & [0] & \dots & [0] \\ [0] & \left[K^2(\mathbf{x}_i^2)\right] & \dots & [0] \\ \vdots & \vdots & \ddots & \vdots \\ [0] & [0] & \dots & \left[K^m(\mathbf{x}_i^m)\right] \end{bmatrix} \qquad \in \mathbb{R}^{m \cdot p \times m \cdot p \cdot (l+u)}$$

and $\mathbf{a}$ is the concatenation of all $\mathbf{a}^m$ into a single column vector

$$\mathbf{a} = \left[\mathbf{a}^1 \dots \mathbf{a}^m\right]^T \qquad \in \mathbb{R}^{m \cdot p \cdot (l+u)}$$

Finally, the concatenation of all the classifications for all samples $\boldsymbol{f}$ can be written in matrix form as:

$$\boldsymbol{K} = \begin{bmatrix} \mathbf{K}(\mathcal{X}_1) \\ \vdots \\ \mathbf{K}(\mathcal{X}_{l+u}) \end{bmatrix} \qquad \in \mathbb{R}^{m \cdot p \cdot (l+u) \times m \cdot p \cdot (l+u)}$$

$$\boldsymbol{f} = \left[\mathbf{f}(\mathcal{X}_1)^T, \dots, \mathbf{f}(\mathcal{X}_{l+u})^T\right]^T = \boldsymbol{K}\mathbf{a} \qquad \in \mathbb{R}^{m \cdot p \cdot (l+u)}$$

Substituting this in Equation 5 yields

$$\left(\frac{1}{l}\mathbf{C}^T\mathbf{C} + \gamma_A I + \gamma_I \mathbf{M}\right)^T \boldsymbol{K}\mathbf{a} = \frac{1}{l}\mathbf{C}^T\mathbf{y}$$

and solving for $\mathbf{a}$

$$\mathbf{a} = \left(\left(\frac{1}{l}\mathbf{C}^T\mathbf{C} + \gamma_A I + \gamma_I \mathbf{M}\right)^T \boldsymbol{K}\right)^{-1} \frac{1}{l}\mathbf{C}^T\mathbf{y} \tag{6}$$

Evaluation on a Test Sample. Once $\mathbf{a}$ is computed, the estimation of the labels/identities of the t probe samples $\mathcal{V} = \{\mathcal{V}_1 \dots \mathcal{V}_t\}$ can proceed ($\mathcal{V}_i$ is a probe sample, analogous to the train samples $\mathcal{X}_i$, and likewise it contains m feature vectors $\mathbf{v}_i^j$ extracted from the ith probe image). First, $\mathbf{f}(\mathcal{V}_i)$ is computed for each image, and the matrix $\boldsymbol{f}(\mathcal{V}) = [\mathbf{f}(\mathcal{V}_1), \dots, \mathbf{f}(\mathcal{V}_t)]^T \in \mathbb{R}^{m \cdot p \cdot t}$ is

composed, with $\mathbf{f}(\mathcal{V}_i)$ the concatenation of all the view classifiers $\mathbf{f}^m$ for probe sample i:

$$\mathbf{f}^m(\mathbf{v}_i^m) = \sum_{j=1}^{l+u} k^m(\mathbf{v}_i^m, \mathbf{x}_j^m)\mathbf{a}_j^m \qquad \in \mathbb{R}^p$$
$$\mathbf{f}(\mathcal{V}_i) = [\mathbf{f}^1(\mathbf{v}_i^1)^T, \ldots, \mathbf{f}^m(\mathbf{v}_i^m)^T]^T \qquad \in \mathbb{R}^{p \cdot m},$$

Let $\boldsymbol{K}(\mathcal{V})$ be the block matrix of the kernels applied to all probe samples:

$$\boldsymbol{K}(\mathcal{V}) = \begin{bmatrix} \mathbf{K}(\mathcal{V}_1) \\ \vdots \\ \mathbf{K}(\mathcal{V}_{l+u}) \end{bmatrix} \qquad \in \mathbb{R}^{m \cdot p \cdot t \times m \cdot p \cdot (l+u)}$$

then $\boldsymbol{f}(\mathcal{V})$ can be directly computed by

$$\boldsymbol{f}(\mathcal{V}) = [\mathbf{f}(\mathcal{V}_1), \ldots, \mathbf{f}(\mathcal{V}_t)]^T = \boldsymbol{K}(\mathcal{V})\mathbf{a} \qquad \in \mathbb{R}^{m \cdot p \cdot t}$$

For the i-th image of the p-th individual, $C \cdot \mathbf{f}(\mathcal{V}_i)$ represents the vector that is as close as possible to $(-1, \ldots, 1, \ldots, -1)$, with 1 at the p-th location. The identity of the i-th image can be estimated *a-posteriori* by taking the index of the maximum value in the vector $C \cdot \mathbf{f}(\mathcal{V}_i)$. In the re-identification field it is customary to output instead of a single identity, a ranked list of possible identities. To create this list the second largest value in the $C \cdot \mathbf{f}(\mathcal{V}_i)$ vector is selected for the second place in the list, and so forth until the p'th place in the list.

During training, the weights $\mathbf{a}$ are learned such as to comply with the labels of the labeled data samples and to promote concordance between classifiers. If training is ran once per test sample, including the test sample in the unlabeled data set, the $\mathbf{a}$ weight vector will be learned once per test sample. These weights will change dynamically during testing: a dynamic classifier.

3.3.1.3 *Kernels*

Any positive definite kernel is a valid choice for use in the multi-view formulation. A few kernels have been tested:

GAUSSIAN: $k(\mathbf{t}, \mathbf{x}) = exp\left(-\frac{\|\mathbf{t}-\mathbf{x}\|^2}{\sigma^2}\right)$

LAPLACIAN: $k(\mathbf{t}, \mathbf{x}) = exp\left(-\frac{|\mathbf{t}-\mathbf{x}|^2}{\sigma^2}\right)$

CHI-SQUARE: $k(\mathbf{t}, \mathbf{x}) = exp\left(-\frac{\sum \frac{(t_i - x_i)^2}{t_i + x_i}}{\sigma^2}\right)$

BHATTACHARYYA: $k(\mathbf{t}, \mathbf{x}) = exp\left(-\frac{\sqrt{1 - \sum \sqrt{t_i \cdot x_i}}}{\sigma^2}\right)$

Any of the above listed kernels can be represented as follows: $k(\mathbf{t},\mathbf{x}) = exp\left(-\frac{D(\mathbf{t},\mathbf{x})}{\sigma^2}\right)$, where $D(.,.)$ is the distance in the numerator of any of the four listed kernels. The respective σ parameter is estimated as $\sigma = \sqrt{2 \cdot D_{med}}$, where D_{med} is the median distance of the distances $D(.,.)$ for all pairs of samples in the training set. This is called a "median estimated kernel bandwidth".

The Chi-Square and Bhattacharrya distances are well suited for comparing histogram features. In the results chapter this is confirmed. For the histogram features used, performance was always better with these kernels.

3.3.2 *Window-based Classifier*

Here the window-based classifier is described. It exploits the temporal coherence of the pedestrian's appearance in the video to increase performance. It takes any single-frame classifier that gives a ranked output, and filters its output.

Instead of providing output for each re-identification, it takes a temporal window and only provides output for a given person if enough re-identifications of a certain rank, of that person are present. This has the effect of filtering out spurious wrong classifications, and recapturing some missed detections and weak (high-rank) re-identifications when they happen between correct strong re-identifications (low rank, lower than a threshold).

In the following, a definition of the main parameters involved in the window-based classifier is provided. These parameters will then be used to tune the operation of a RE-ID system:

- **Rank (r):** Given an ordered list of the matching scores (sorted in descending order) of a probe sample against all gallery samples, rank denotes the largest index in the ordered list in which the correct match for that sample may show up (see illustration in Figure 20). It can also be used as a sensitivity parameter to set the algorithm operating point (*e.g.,* accepting re-identifications of high rank will improve recall and decrease precision).
- **Window size (w):** This stands for the number of frames of the window under consideration.
- **Detection threshold (d):** This variable controls the required minimum number of re-identifications of rank r in a window of size w frames for that window to be considered a positive re-identification window.

Therefore, a window of w frames is considered a positive detection of a certain person if it has at least d detections whose respective re-identifications of that person are of rank r. Intuitively, for larger r a lot more re-identifications will be accepted and thus less precision will ensue, but likely more recall. For larger d, more concordant re-identifications are required to give output, less output will be given, therefore precision will likely increase while recall

Figure 20: Explanation of Rank. Matteo appears in the video and is detected twice. He is matched against all pedestrians in the gallery set, and the classifier outputs an ordered list for each detection. Considering Rank1 re-identifications to signal a positive re-identification, Matteo is only correctly re-identified once. Considering up to Rank3 Matteo is correctly re-identified in both frames.

diminishes. Finally for larger w, there will be more chances for the requested d re-identifications of rank r to be captured and thus recall will likely increase at a cost of precision since so much more output in the form of video-size will be given (confirmation of this intuition is given in Table 15).

Based on the parameters just introduced above, I propose using such triplet of parameters $\mathcal{T} = (r, d, w)$, to tune the algorithm's performance. For instance one detection with a corresponding re-identification of Rank 1 (d=1 and r=1) usually does not provide enough/reasonable confidence to justify giving output to a human operator. In fact, given the low rank 1 re-identification rate of the RE-ID algorithms in the literature (around 30%), it's required to have several rank 1 re-identifications of a pedestrian, in a short period of time, to have a reasonable confidence that the pedestrian is indeed present. Therefore, I studied the necessary rank r, size of window w and required number of detections d to optimize performance of the tested classification algorithms, and defined guidelines on how to change these parameters to improve some particular aspects, *e.g.*, precision vs recall (see Table 15 and Section 4.4.8 for the results and discussion on this matter).

Although this procedure is similar to multi-shot (see Section 1.3), it requires less information. In this work, window-based classification works with any single-shot re-identification algorithm, and does not require an in-camera tracker, contrary to the majority of the works that do multi-shot re-identification.

3.3.3 *Clip-based Output*

Video-clips that encapsulate frames with detections and RE-IDs of the person of interest will be used as the output of the system in order to decrease the attentional load of the user.

This proposal is supported by the following four reasons: *(a)* A single detection and respective re-identification does not guarantee a high degree confidence, therefore several of them are desirable to have higher confidence; *(b)* Browsing a sequence frame by frame takes significantly more time than observing a video with the same number of frames; *(c)* Pedestrian appearance in frames is not independent, they almost always appear in several contiguous frames. *(d)* The presence of motion traits in videos helps human operators recognize and validate the re-identified pedestrians. Therefore, providing output in the form of video-clips, encapsulating several positive detections and RE-IDs of one given person, is well suited to address the above concerns.

One video-clip is generated for the union of all positive windows that overlap or are contiguous. In Figure 14 I show an example with window size equal to 4 frames (w=4), minimum number of detections of 2 (d=2) and rank one (r=1). The person appears in 4 frames and is only detected and re-identified in two frames (the only two red bounding boxes). Note how, albeit only being re-identified in 2 frames, the final output video-clip contains all 4 frames of interest.

3.4 INTER-CAMERA TRACKING

State-of-the-art re-identification algorithms have poor rank 1 classification rates (~30%). To raise the overall performance the re-identification stage is integrated into a over-arching inter-camera tracking system that employs the Multiple Hypothesis algorithm [6]. Adding the temporal dimension to the problem, plus spatial constraints to the movement of pedestrians in the system, makes the problem more tractable. Also adding this algorithm's ability to correct spurious mis-classifications of the past, makes the overall system even able to disambiguate cases where pedestrians partially change their attire [6].

The Multiple Hypothesis Tracking (MHT) algorithm was adopted to implement the inter-camera tracking ability. This algorithm keeps multiple interpretations of the current persons' locations in the camera network using both temporal and spacial constraints, taking into account the topology of the camera network and the connectivity of the space. For instance, if a person was detected in a certain camera, the likelihood that it is found in neighbor locations at neighbor times increases. This disambiguation capability allows for the resolution of past mis-associations when more information is available. The granularity of the detections is defined by coarse zones, usually one zone for each camera. In other words, inter-camera tracking is done in a graph and thus does not require precise incremental-locations, as needed for example with (x, y) tracking in the field of view of a single camera. With a coarse reso-

lution for tracking, missed detections are more tolerable, allowing the detector to be tuned to significantly reduce false positives.

3.4.1 *Multiple Hypothesis Tracking algorithm*

In its original formulation, the MHT algorithm is used to track various targets over two or three dimensional spaces [107]. The algorithm continuously maintains a set of hypotheses on the various possible states of the world. Each hypothesis contains information on the existing targets, and their tracks. Each has a probability of being correct. The system periodically receives new scans containing data from the sensors. All the measurements in time k are denoted by Z^k, and the measurement l of time k is denoted by Z_l^k. Each measurement corresponds to an observation, and is usually associated with a (x, y) or (x, y, z) position in space and possibly other additional target features, such as target size. Let Ω_i^k denote the hypothesis i in scan k. Each hypothesis Ω_i^k contains a set T_i^k of existing targets ${}^{\iota}T_i^k$ ($\iota \in [1, ..., n]$ targets), the state estimate for each target, the state estimate covariance, and the association ψ_i^k, between the measurements Z^k and the hypothesized targets T_i^k. Every hypothesis Ω_i^k is associated with a probability p_i^k.

At each time instant k, the hypotheses Ω^{k-1} are used to produce the hypotheses Ω^k. For each hypothesis Ω_j^{k-1} a new set of hypotheses is generated ${}^{j}\Omega^k$ which have Ω_j^{k-1} as parent (superscript j indicates hypothesis with parent j). In the generation of the new set of hypotheses ${}^{j}\Omega^k$, each observation Z_l^k is considered to be either a False Positive (FP), a New Target (NT), or a detection of an existing target. However, an observation Z_l^k is only considered to have origin in a target ${}^{\iota}T_i^k$ of hypothesis Ω_j^{k-1} if it falls in the target's gate (area around target's expected position) – which is calculated based on the covariance of the state estimate. Furthermore, often each observation can only be assigned to at most one target, and each target can only be assigned to at most one observation (group tracking is addressed by Mucientes and Burgard [97]). A target track is terminated if the target is not detected after t time steps.

The probability of a new hypothesis ${}^{j}\Omega_i^k$ given the parent hypothesis Ω_j^{k-1} and the measurements Z^k, is

$$
\begin{aligned}
{}^{j}p_i^k = & \frac{1}{c} \times {P_d}^{N_d} \times (1 - P_d)^{N_t - N_d} \times (P_{FP})^{N_{fp}} \times \\
& (P_{NT})^{N_{nt}} \times \prod_{(Z_l^k, {}^{\iota}T_i^k) \in \psi_i^k} P_{Z_l^k, {}^{\iota}T_i^k} \times p_j^{k-1}
\end{aligned} \tag{7}
$$

where N_d corresponds to the number of measurements and N_t to the number of targets in Ω_j^{k-1}, N_{fp} is the number of false positives and N_{nt} is the number of new targets [107]. Furthermore, P_d is the probability of detecting a target, P_{FP} the probability of a measurement being a false positive, and P_{NT} the probably of detecting a new target. The probability of the parent hypothesis is p_j^{k-1}, and $P_{Z_l^k, {}^{\iota}T_i^k}$ denotes the probability that measurement Z_l^k is a detection

of target ${}^{\iota}T_i^k$, which is usually calculated based on the target position estimate, and the covariance of this estimate.

The algorithm generates a combinatorial explosion of hypotheses. This exponential growth of the number of hypotheses can be controlled by pruning the hypotheses tree. Usual pruning strategies include limiting the number of leaves, or the depth of the tree [15]. However, while generating the hypotheses ${}^{j}\Omega^k$, for a single leaf (Ω_j^{k-1}), the number of hypotheses to generate can be too large to process in real time. For example, if there were 30 targets in Ω_j^{k-1}, and Z^k contains 30 measurements there will be 6.2×10^{37} hypotheses in ${}^{j}\Omega^k$ (for more details on calculating the number of generated hypotheses see Danchick and Newnam [32]). These hypotheses will eventually be pruned, after the hypotheses for all leaves are generated, but the processing time and memory space that the explicit enumeration of all these hypotheses consumes is insupportable. A solution is to use an algorithm due to Murty to find the ranked k-best assignments for the association in each leaf [30], instead of explicitly enumerating all the possible hypotheses. Clustering, which consists of dividing the hypotheses tree into several trees taking advantage of the independence between the tracks of some targets, can also be used to reduce the processing requirements of MHT and increase its performance [107].

To implement the MHT algorithm, the Multiple Hypothesis Library was used, described by Antunes *et al.* in [5]. This library already handles clustering, and provides pruning of the tree limiting both the tree depth and the number of leaves. The Murty algorithm for finding the k-best assignments is also implemented.

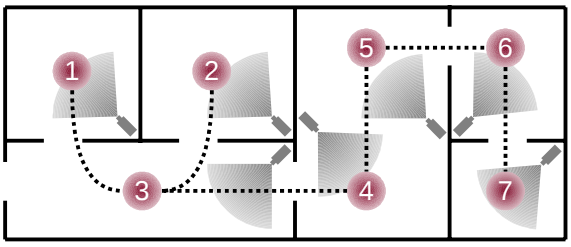
Below it is described the application of the MHT algorithm to the specific problem of tracking on a multi-camera network with non-overlapping fields of view, which is the most common case in video surveillance systems.

3.4.1.1 *Graph representation*

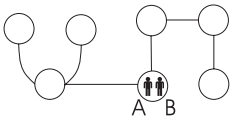
Let the tracking area be represented as a graph. Let $G = (A, C)$ denote the graph representing the tracking area, where A consists of a set of tracking zones $A = \{z_1, ..., z_n\}$ and C of a set of connections between zones. Thus, (z_i, z_j) is an edge belonging to C if and only if z_i and z_j have a connection [99]. The topology of the graph can be manually defined or learned automatically [54].

For our particular problem, each zone is associated with one camera, and each camera is associated with one zone. Even though it is possible to divide the field of view of a camera into different zones, which may be useful in some specific situations, this possibility is not addressed in this work.

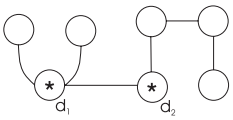
A possible scenario is presented in Figure 21 (a). Several cameras are spread throughout the tracking area, and each camera monitors a division or part of a division. The circles represent the zones in the graph and the dotted lines the connections between them.



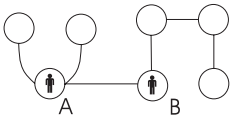
(a) Floor map, cameras and zones graph



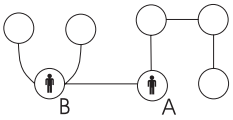
(b) Initial poses



(c) Detections



(e) Hypothesis 1



(f) Hypothesis 2

Figure 21: Example of a tracking area and the zones graph. Each camera has a field of view (gray area) which defines a single zone (a). Given an initial configuration where two persons, A and B, are in zone 4 (b), and then two target detections occur in zones 3 and 4 (c), one has various possibilities of localization of the two targets. Assuming both detections are valid, and related to the targets A and B, then one has two hypothesis, A in 3 and B in 4 (d), or vice versa, B is in 3 and A is in 4.

Because the tracking area is a graph, each detection $Z_l^k \in Z^k$ is associated with a zone $z \in A$, instead of (x, y) coordinates. Each detection also contains a set of features which describe the detected target, which will now be discussed.

3.4.1.2 *Tracking granularity*

In the proposed approach, targets are tracked across multiple cameras, and not locally, in the (x, y) field of view of each single camera. It would also be possible to perform the tracking of the (x, y) position of targets in each camera, which is the usual case for tracking.

Contrary to the fine (x, y) tracking, when tracking across zones, the requirements on the pedestrian detection performance can be reduced. This makes it possible to use tighter thresholds for detection reducing the number of false positives, but also the number of true positives as well. This would not be possible if tracking was done in the field of view of a single camera, as the reduction of true positives would result in many lost tracks. On the contrary,

when tracking across cameras, it is not as necessary to have many detections of the same target, in the same zone, in sequence.

There are some particular situations where finer grained tracking is necessary. This may happen with cameras covering a large field of view, with high resolution and several small targets. In this case, the field of view of the camera may be divided into a grid of separate zones, in which case the proposed solution is directly applicable. Furthermore, local tracking in each camera can always be performed if necessary, in parallel with the proposed approach.

3.4.1.3 *Integration with the MHT Algorithm*

Each detection Z_l^k contains the state information about each target, which includes the target identifier, the zone where the target is, the features that describe the detected person, and the time of the target's last detection.

The probability $P_{Z_l^k, \iota T_i^k}$ of measurement Z_l^k being a detection of target ιT_i^k is calculated taking into consideration the zone where the target was, the one where the measurement is taken, the features associated with the target, and the ones associated with the measurement. For a detection Z_l^k and a target ιT_i^k, let h^Z be the feature histograms associated with the detection, and h^T the feature histograms associated with the target. Also, let z_D and z_T be respectively the zone associated with the detection and the zone where the target was in the hypothesis Ω_j^{k-1}.

The probability $P_{Z_l^k, \iota T_i^k}$ is calculated as:

$$P_{Z_l^k, \iota T_i^k} = P_{h^Z, h^T} \cdot P_{z_D, z_T} \tag{8}$$

The probability P_{h^Z, h^T} depends on the difference between the histograms, which can be calculated using the Hellinger's distance:

$$B(h^Z, h^T) = \sqrt{1 - \sum_{i=1}^{m} \sqrt{(h_i^Z \cdot h_i^T)}} \tag{9}$$

where m is the number of bins in the color histograms. The Hellinger's distance is then used to calculate P_{h^Z, h^T}:

$$P_{h^Z, h^T} = \left(1 + \lambda \cdot B(h^Z, h^T)\right)^{-1} \tag{10}$$

The probability P_{h^Z, h^T} will be in the interval $[\frac{1}{\lambda+1}, 1]$. The value of λ should be chosen to obtain the desired minimum value for probability P_{h^Z, h^T}.

The probability P_{z_D, z_T} is 1 when $z_D = z_T$. For other cases, there are several manners in which P_{z_D, z_T} can be calculated. In the simplest form, $P_{z_D, z_T} = c$, where c is a constant probability of transition between zones, when $(z_D, z_T) \in C$ (the zones have a connection), and $P_{z_D, z_T} = 0$ when $(z_D, z_T) \notin C$. A more flexible approach includes a probability transition matrix, M, such that M_{z_i, z_j} contains the probability of transition between zones z_i and z_j, then $P_{z_D, z_T} = M_{z_D, z_T}$ when $(z_D, z_T) \in C$. Gilbert and Bowden provide a method for the automatic learning of M [54].

The most complex case occurs when $(z_D, z_T) \notin C$ and $P_{z_D,z_T} \neq 0$ is required, that is, the target is detected in a zone which does not have a direct connection with the one in which it was before, and a probability modeling which does not simply assign 0 to P_{z_D,z_T} is required. In this case, the person crossed one or more zones without being detected. Therefore, there is not a single path that he could have taken from z_T to z_D, but many possible paths. Because it is impossible to determine exactly which of the possible paths was taken, and no future information will help with this task, the path with the greatest probability of being the correct one should be chosen. This path will naturally correspond to the one that maximizes the product of the probability of transition between all the zones in the path. This is the problem of finding the shortest path in a graph, and is usually solved using the Dijkstra algorithm. Fortunately, because the matrix M is constant over time, the shortest paths between all the zones in the graph can be precomputed using the Floyd-Warshall algorithm [28].

3.4.1.4 *Entry zone*

When the tracking area of interest is in the interior of a closed building or sealed area it is possible to greatly improve the tracking results by defining one or more entry/exit zones. In a closed building, new targets cannot appear in all the tracking zones. Usually, there are a few entrances where the targets can enter and leave the tracking area, which is the case with the example in Figure 21. In the tracking area represented in the figure, a target track can only initiate and terminate in zone 3. If this information is included in the tracker, then detections in every other zone will only be attributed to either false positives or existing targets, and targets in those zones will not be deleted, even if they are not detected after a long period of time.

Algorithm 1 Multiple Hypothesis algorithm

procedure MAIN
  $P_d \leftarrow$ prior for re-identification
  $P_{NT} \leftarrow$ prior for new targets
  $P_{FP} \leftarrow$ prior for false positives
  Hypothesis set $\Omega^0 \leftarrow$ empty
Notation:
  Re-identifications set in time k : Z^k
  Re-identification l *of time* k : Z_l^k
  Hypothesis i *in time* k : Ω_i^k
hypothesis Ω_i^k *contains:*
    set T_i^k *of existing targets* ${}^{\iota}T_i^k$ *(*$\iota \in [1, ..., n]$ *targets),*
    state estimate for each target,
    state estimate covariance,
    association ψ_i^k*, between RE-IDs* Z^k *and hypothesized targets* T_i^k
    probability of self p_i^k*.*
  loop
    $Z^k \leftarrow$ re-identifications
    if Only one re-identification in Z^k **then**
      for all hypothesis Ω_i^{k-1} **do**
        Call algorithm by Murty [30] to find k-best hypothesis instead of enumerating all possible hypotheses below
        if RE-ID Z_1^k in entry zone **then** ▷ *new target*
          Create ${}^{i}\Omega_j^k$ from Ω_i^{k-1} with added target
        if RE-ID Z_1^k not in entry zone **then** ▷ *new target with missed detections before*
          Create ${}^{i}\Omega_j^k$ from Ω_i^{k-1} with added target
          p_j^k given by shortest path between the entry zone and current zone of target (path that maximizes transition probability between all the zones in the path)
        if RE-ID Z_1^k not in gate of any T_i^k **then**
          Create ${}^{i}\Omega_j^k$ from Ω_i^{k-1} with no change ▷ *FP*
          for all existing targets T_i^k **do** ▷ *positive re-identification outside gate (missed detection before)*
            Create ${}^{i}\Omega_j^k$ with updated location of T_i^k
            p_j^k given by shortest path between the two zones (path that maximizes transition probability between the previous zone of target and the current zone in the path)
        if RE-ID Z_1^k in gate of at least one T_i^k in Ω_i^{k-1} **then**
          Create ${}^{i}\Omega_j^k$ from Ω_i^{k-1} with no change ▷ *FP*
          for all existing targets T_i^k **do** ▷ *positiveRE-IDs*
            Create ${}^{i}\Omega_j^k$ with updated location of T_i^k
        for all existing targets T_i^k **do**
          if Target T_i^k not detected in n time steps **then**
            Delete target
    elseif More than one re-identification in Z^k **then**
      Create hypothesis for all the combinations of the above enumerated cases
    Prune low probability hypotheses

4 RESULTS

In this chapter we go over all the relevant results obtained. First the benefit of the feature extraction process is evaluated in Section 4.1. Then the performance of the Multi-View classifier is assessed in Section 4.2. Section 4.3 illustrates an example of the MHT algorithm in action. By the end of the chapter in Section 4.4, results on the integration between Pedestrian Detection (PD) and Re-Identification (RE-ID) are put forth.

4.1 DESCRIPTOR EXTRACTION COMPARISON

In this section, standard re-identification experiments were run. Standard re-identification experiments consider manually segmented pedestrians, re-identification in single frames and a closed space scenario (all persons detected are in the gallery) and short-term time span (persons do not change clothes). This to illustrate the benefits of the proposed descriptor extraction method. These initial experiments were run in three datasets, with varying combinations of features, use of equalization on the features, and NN classifiers, for each of the four descriptor extraction methods.

4.1.1 *Features used*

The features employed were:

- Hue-Saturation-Value histogram (HSV)[59];
- Black-Value-Tint histogram (BVT) is a variant of HSV developed for [25]. It is constructed as follows: First, count all the black and near-black[1] pixels (where the Hue and Saturation values are basically random) and attribute them to one bin (the B of BVT, for Black pixels). Then for the rest of the non-black pixels, make (1) a regular Value (gray-scale) histogram vector (the V of BVT, for Value histogram), and (2) a 2D histogram matrix from the Hue and Saturation values (the T of BVT, for Tint histogram.).
- Lightness color-opponent histogram (Lab)[64];
- Maximum Response Filter Bank (MR8) histogram [75, 109];
- Local Binary Patterns (LBP) histogram [2].

Each feature when applied to a region of the image generates an histogram of constant bin size for all experiments (illustrated in Figure 22).

1 Definition of "near-black": The value/grey-scale channel of the image is equalized and quantized into ten bins. The "near-black" pixels are those that fall into the darkest bin of the ten.

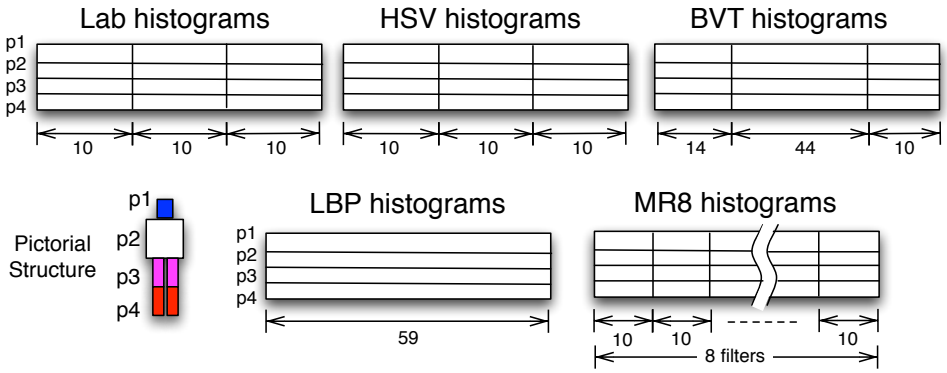


Figure 22: Different features represented by blocks are computed from the detected parts $\{p_i\}_{i=1}^{4}$.

4.1.2 *Classifiers used*

In these experiments more complex classifiers are not used because the objective is to test the descriptor classifiers only. The NN classifiers employed used the following distances:

- **Bhatt** Hellinger's distance: $D(x,t) = \sqrt{1 - \sum_{i=1}^{d} \sqrt{x_i \cdot t_i}}$,
- **ChiSq** Chi-Squared distance: $D(x,t) = \sum_{i=1}^{d} \frac{(t_i - x_i)^2}{t_i + x_i}$,
- **Diffusion** Diffusion distance [80],
- **Euclidean** Euclidean distance: $D(x,t) = \sqrt{\sum_{i=1}^{d} x_i \cdot t_i}$

where x and t are normalized feature vectors of size d, obtained from the concatenation of the several histograms represented in Figure 22. *I.e.*, for HSV, $d = 120$. When using these NN classifiers, (1st) features are extracted from all images, (2nd) a distance matrix all-to-all is computed, and (3rd) the minimum distance from each probe to all gallery images if found, to determine the nearest-neighbor match for each probe image.

4.1.3 *Datasets used*

In these experiments the VIPeR, iLIDS4REID and 3DPeS datasets were used (sample images in Figures 23, 24 and 25). These are well established datasets used by most re-identification works in the literature. VIPeR contains 632 pairs of 128×64 images of 632 pedestrians, captured from two different cameras. iLIDS4REID contains 476 images of 119 pedestrians, captured from up to two different cameras inside an airport. 3DPeS contains 605 images of 199

pedestrians, captured from up to eight different cameras on a college campus environment.

For each experiment in each dataset, 100 runs were made, and the results shown are the average of those 100 runs. For each run in the VIPeR dataset, 316 pedestrians were randomly selected, and one image of the pair was taken at random to be the probe and the other to be in the gallery. For each run in the iLIDS4REID or 3DPeS datasets, two images were randomly selected from each pedestrian, one to be the probe and the other to be in the gallery.

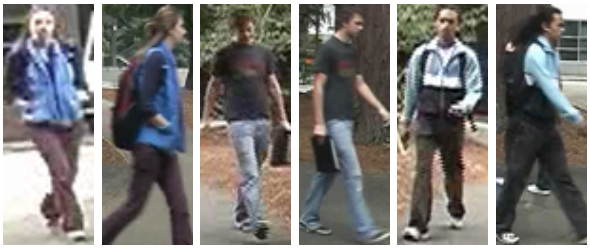


Figure 23: Sample images from the VIPeR dataset. It has only two images for each pedestrian, from two distinct cameras, in an outdoors environment. Almost all pairs have the respective pedestrian in different poses, facing different directions with about a 90° different angle.

Figure 24: Sample images from the iLIDS4REID dataset. It contains a few images of each pedestrian from up to two different camera views in an airport.

Figure 25: Sample images from the 3DPeS dataset. It contains some images from up to eight different camera views in a college campus environment.

Dataset	Feature	Equalization	Descriptor Extraction	NN Classifier	Rank1 (%)	Rank5 (%)	nAUC (%)
VIPeR	MR8	×	1 Part	Euclidean	01.5	03.7	53.7
VIPeR	MR8	×	2 Parts	Euclidean	01.3	04.4	61.5
VIPeR	MR8	×	6 Parts	Euclidean	01.6	05.8	61.2
VIPeR	MR8	×	4 Parts (proposed)	Euclidean	01.4	05.8	61.5
VIPeR	MR8	×	1 Part	Diffusion	00.8	04.0	53.1
VIPeR	MR8	×	2 Parts	Diffusion	01.6	04.4	60.6
VIPeR	MR8	×	6 Parts	Diffusion	01.6	06.4	60.5
VIPeR	MR8	×	4 Parts (proposed)	Diffusion	01.8	06.6	61.3
VIPeR	MR8	×	1 Part	ChiSq	00.9	04.2	53.7
VIPeR	MR8	×	2 Parts	ChiSq	01.8	05.0	61.7
VIPeR	MR8	×	6 Parts	ChiSq	01.3	06.2	61.3
VIPeR	MR8	×	4 Parts (proposed)	ChiSq	02.0	06.5	62.1
VIPeR	MR8	×	1 Part	Bhatt	00.8	03.8	54.0
VIPeR	MR8	×	2 Parts	Bhatt	01.9	04.8	62.2
VIPeR	MR8	×	6 Parts	Bhatt	01.5	06.2	62.0
VIPeR	MR8	×	4 Parts (proposed)	Bhatt	02.0	06.6	62.7
VIPeR	Lab	×	1 Part	Euclidean	05.2	13.2	73.0
VIPeR	Lab	×	2 Parts	Euclidean	09.5	21.1	77.4
VIPeR	Lab	×	6 Parts	Euclidean	10.5	21.9	78.1
VIPeR	Lab	×	4 Parts (proposed)	Euclidean	11.2	21.8	78.2
VIPeR	Lab	×	1 Part	Diffusion	05.3	13.4	73.5
VIPeR	Lab	×	2 Parts	Diffusion	10.6	22.7	78.8
VIPeR	Lab	×	6 Parts	Diffusion	11.5	24.4	80.0
VIPeR	Lab	×	4 Parts (proposed)	Diffusion	12.0	25.4	79.8
VIPeR	Lab	×	1 Part	ChiSq	06.6	16.4	74.3
VIPeR	Lab	×	2 Parts	ChiSq	13.1	25.2	80.0
VIPeR	Lab	×	6 Parts	ChiSq	12.9	26.3	80.7
VIPeR	Lab	×	4 Parts (proposed)	ChiSq	13.0	27.2	80.8
VIPeR	Lab	×	1 Part	Bhatt	07.2	16.8	74.3
VIPeR	Lab	×	2 Parts	Bhatt	13.6	26.9	80.2
VIPeR	Lab	×	6 Parts	Bhatt	13.1	26.8	80.7
VIPeR	Lab	×	4 Parts (proposed)	Bhatt	13.6	28.6	80.8

Table 4: Results in the VIPeR dataset, for the MR8 and Lab features, with the Bhaat, Chisq, Diffusion and Euclidean distances in the NN classifier. The best results for the descriptor extraction method for each feature and classifier combination is shown underlined. Equalization indicates if histogram equalization was applied to the dataset or not.

4.1.4 *Results*

For these experiments the standard RE-ID metric, the Cumulative Matching Characteristic curve (CMC) was used. In Tables 4, 5, 6 and 7 it is reported the first rank percentage, the fifth rank percentage and the normalized area under the CMC. The results are coherent across almost all datasets, features and NN classifiers tested. Dividing the body in 4 parts [head | torso | thighs | forelegs] (as shown in Figure 18d) for descriptor extraction outperforms in almost all cases the other descriptor extraction methods. Also, detecting the 6 body parts and treating them separately for the purpose of descriptor extraction consistently surpasses just dividing the body in two parts (above-waist and below-waist). Finally, the waist-division descriptor extraction method consistently exceeds extracting features from the whole-body.

Dataset	Feature	Equalization	Descriptor Extraction	NN Classifier	Rank1 (%)	Rank5 (%)	nAUC (%)
VIPeR	HSV	×	1 Part	Euclidean	06.1	15.9	75.9
VIPeR	HSV	×	2 Parts	Euclidean	10.8	23.7	80.5
VIPeR	HSV	×	6 Parts	Euclidean	11.5	25.5	81.4
VIPeR	HSV	×	4 Parts (proposed)	Euclidean	11.5	26.5	81.2
VIPeR	HSV	×	1 Part	Diffusion	06.6	15.8	75.9
VIPeR	HSV	×	2 Parts	Diffusion	11.3	25.6	81.6
VIPeR	HSV	×	6 Parts	Diffusion	13.9	26.7	82.3
VIPeR	HSV	×	4 Parts (proposed)	Diffusion	14.0	27.4	82.5
VIPeR	HSV	×	1 Part	ChiSq	07.0	17.6	77.3
VIPeR	HSV	×	2 Parts	ChiSq	12.3	27.5	82.5
VIPeR	HSV	×	6 Parts	ChiSq	13.6	29.5	83.1
VIPeR	HSV	×	4 Parts (proposed)	ChiSq	14.3	30.6	83.2
VIPeR	HSV	×	1 Part	Bhatt	07.4	17.8	77.7
VIPeR	HSV	×	2 Parts	Bhatt	12.4	27.4	83.1
VIPeR	HSV	×	6 Parts	Bhatt	13.3	29.3	83.4
VIPeR	HSV	×	4 Parts (proposed)	Bhatt	14.5	31.2	83.7
VIPeR	BVT	×	1 Part	Euclidean	06.9	16.5	76.0
VIPeR	BVT	×	2 Parts	Euclidean	09.3	22.6	80.4
VIPeR	BVT	×	6 Parts	Euclidean	11.3	23.1	77.0
VIPeR	BVT	×	4 Parts (proposed)	Euclidean	12.1	24.6	79.0
VIPeR	BVT	×	1 Part	Diffusion	07.6	18.2	77.6
VIPeR	BVT	×	2 Parts	Diffusion	13.0	27.7	82.7
VIPeR	BVT	×	6 Parts	Diffusion	14.6	29.9	83.3
VIPeR	BVT	×	4 Parts (proposed)	Diffusion	15.0	30.5	83.4
VIPeR	BVT	×	1 Part	ChiSq	09.0	21.1	79.1
VIPeR	BVT	×	2 Parts	ChiSq	14.6	31.8	83.7
VIPeR	BVT	×	6 Parts	ChiSq	16.3	35.4	84.1
VIPeR	BVT	×	4 Parts (proposed)	ChiSq	17.2	36.5	84.1
VIPeR	BVT	×	1 Part	Bhatt	09.3	22.1	79.4
VIPeR	BVT	×	2 Parts	Bhatt	15.2	32.7	84.1
VIPeR	BVT	×	6 Parts	Bhatt	17.3	35.7	84.4
VIPeR	BVT	×	4 Parts (proposed)	Bhatt	17.9	36.9	84.5

Table 5: Results in the VIPeR dataset, for the BVT and HSV features, with Bhaat, Chisq, Diffusion and Euclidean distances in the NN classifier. The best results for the descriptor extraction method for each feature and classifier combination is shown underlined. Equalization indicates if histogram equalization was applied to the dataset or not.

Another observable result is how BVT almost always outperforms the other tested features, and how the Hellinger's distance always beats the other tested NN classifiers, all other factors the same.

4.1.5 *Discussion*

As expected, dividing the body in two parts (below the waist and above the waist) provides more information and thus more discriminatory power over extracting descriptors from the whole body regardless of body-location. Dividing the body further into six parts [head | torso | thigh | thigh | fore-leg | fore-leg] further increases the resolution of the description extraction, thus increasing the discriminatory power. By allowing the description extraction to treat the head separately from the torso, and the shins separately from the

Dataset	Feature	Equalization	Descriptor Extraction	NN Classifier	Rank1 (%)	Rank5 (%)	nAUC (%)
3DPeS	HSV	×	1 Part	Diffusion	11.4	25.0	81.0
3DPeS	HSV	×	2 Parts	Diffusion	18.7	39.4	85.8
3DPeS	HSV	×	6 Parts	Diffusion	21.9	43.2	87.0
3DPeS	HSV	×	4 Parts (proposed)	Diffusion	22.4	44.4	87.1
3DPeS	HSV	✓	1 Part	Diffusion	07.3	21.0	76.3
3DPeS	HSV	✓	2 Parts	Diffusion	15.0	34.1	84.1
3DPeS	HSV	✓	6 Parts	Diffusion	18.9	39.0	86.4
3DPeS	HSV	✓	4 Parts (proposed)	Diffusion	19.7	39.6	86.7
3DPeS	HSV	×	1 Part	Euclidean	11.4	27.1	79.2
3DPeS	HSV	×	2 Parts	Euclidean	17.4	37.9	84.2
3DPeS	HSV	×	6 Parts	Euclidean	20.5	41.9	86.0
3DPeS	HSV	×	4 Parts (proposed)	Euclidean	20.9	42.1	86.1
3DPeS	HSV	✓	1 Part	Euclidean	07.9	21.7	76.5
3DPeS	HSV	✓	2 Parts	Euclidean	15.7	34.4	83.9
3DPeS	HSV	✓	6 Parts	Euclidean	17.1	39.9	86.3
3DPeS	HSV	✓	4 Parts (proposed)	Euclidean	18.0	40.4	86.6
3DPeS	HSV	×	1 Part	ChiSq	12.7	28.3	81.3
3DPeS	HSV	×	2 Parts	ChiSq	19.8	41.4	85.8
3DPeS	HSV	×	6 Parts	ChiSq	22.9	44.0	86.9
3DPeS	HSV	×	4 Parts (proposed)	ChiSq	23.7	45.7	87.3
3DPeS	HSV	✓	1 Part	ChiSq	07.7	21.8	76.0
3DPeS	HSV	✓	2 Parts	ChiSq	15.4	35.0	84.4
3DPeS	HSV	✓	6 Parts	ChiSq	19.2	39.6	86.8
3DPeS	HSV	✓	4 Parts (proposed)	ChiSq	19.5	40.3	87.0
3DPeS	HSV	×	1 Part	Bhatt	12.7	28.2	81.2
3DPeS	HSV	×	2 Parts	Bhatt	21.0	43.4	85.5
3DPeS	HSV	×	6 Parts	Bhatt	23.4	45.2	86.7
3DPeS	HSV	×	4 Parts (proposed)	Bhatt	24.7	48.1	87.2
3DPeS	HSV	✓	1 Part	Bhatt	07.9	21.7	75.7
3DPeS	HSV	✓	2 Parts	Bhatt	16.1	34.4	83.7
3DPeS	HSV	✓	6 Parts	Bhatt	19.3	39.3	86.1
3DPeS	HSV	✓	4 Parts (proposed)	Bhatt	18.7	41.0	86.4
3DPeS	BVT	×	1 Part	Bhatt	16.5	33.2	81.5
3DPeS	BVT	×	2 Parts	Bhatt	22.4	44.2	85.5
3DPeS	BVT	×	6 Parts	Bhatt	25.1	46.5	87.1
3DPeS	BVT	×	4 Parts (proposed)	Bhatt	26.3	46.7	87.6

Table 6: Results in the 3DPeS dataset, with the HSV feature, applying histogram equalization to the dataset's images or not, for all four distances in the NN classifier. Results show that BVT is the best single feature overall. The best results for the descriptor extraction method for each feature and classifier combination is shown underlined. Equalization indicates if histogram equalization was applied to the dataset or not.

Dataset	Feature	Equalization	Descriptor Extraction	NN Classifier	Rank1 (%)	Rank5 (%)	nAUC (%)
iLIDS4REID	MR8	×	1 Part	Bhatt	05.7	20.8	70.9
iLIDS4REID	MR8	×	2 Parts	Bhatt	09.2	24.5	72.0
iLIDS4REID	MR8	×	6 Parts	Bhatt	08.8	25.0	73.7
iLIDS4REID	MR8	×	4 Parts (proposed)	Bhatt	09.9	25.1	74.4
iLIDS4REID	MR8	✓	1 Part	Bhatt	06.4	17.9	68.4
iLIDS4REID	MR8	✓	2 Parts	Bhatt	10.1	26.4	73.2
iLIDS4REID	MR8	✓	6 Parts	Bhatt	12.3	28.7	75.9
iLIDS4REID	MR8	✓	4 Parts (proposed)	Bhatt	12.9	31.3	76.8
iLIDS4REID	Lab	×	1 Part	Bhatt	14.3	32.8	78.0
iLIDS4REID	Lab	×	2 Parts	Bhatt	20.3	38.6	81.2
iLIDS4REID	Lab	×	6 Parts	Bhatt	21.8	44.8	82.3
iLIDS4REID	Lab	×	4 Parts (proposed)	Bhatt	22.4	45.1	83.1
iLIDS4REID	Lab	✓	1 Part	Bhatt	09.8	21.9	73.0
iLIDS4REID	Lab	✓	2 Parts	Bhatt	14.5	31.6	79.5
iLIDS4REID	Lab	✓	6 Parts	Bhatt	18.2	38.5	81.9
iLIDS4REID	Lab	✓	4 Parts (proposed)	Bhatt	18.5	39.1	83.0
iLIDS4REID	HSV	×	1 Part	Bhatt	13.9	31.9	77.1
iLIDS4REID	HSV	×	2 Parts	Bhatt	20.0	37.7	80.9
iLIDS4REID	HSV	×	6 Parts	Bhatt	22.2	42.2	81.9
iLIDS4REID	HSV	×	4 Parts (proposed)	Bhatt	22.3	44.6	82.7
iLIDS4REID	HSV	✓	1 Part	Bhatt	09.7	25.6	74.2
iLIDS4REID	HSV	✓	2 Parts	Bhatt	15.4	31.3	80.0
iLIDS4REID	HSV	✓	6 Parts	Bhatt	19.3	39.2	81.8
iLIDS4REID	HSV	✓	4 Parts (proposed)	Bhatt	19.8	39.3	82.9
iLIDS4REID	BVT	×	1 Part	Bhatt	22.2	43.8	80.1
iLIDS4REID	BVT	×	2 Parts	Bhatt	21.3	40.0	80.7
iLIDS4REID	BVT	×	6 Parts	Bhatt	25.5	48.2	85.4
iLIDS4REID	BVT	×	4 Parts (proposed)	Bhatt	25.8	49.5	85.3
iLIDS4REID	BVT	✓	1 Part	Bhatt	10.3	22.1	71.2
iLIDS4REID	BVT	✓	2 Parts	Bhatt	14.3	28.8	76.2
iLIDS4REID	BVT	✓	6 Parts	Bhatt	16.8	32.5	80.0
iLIDS4REID	BVT	✓	4 Parts (proposed)	Bhatt	17.0	34.5	80.7

Table 7: Results in the iLIDS4REID dataset, with the BVT, HSV, Lab and MR8 features, applying histogram equalization to each dataset or not, for the Bhatt distance in the NN classifier. The best results for the descriptor extraction method for each feature and classifier combination is shown underlined. Equalization indicates if histogram equalization was applied to the dataset or not.

thighs, this enables features to be extracted from more local regions in the persons body.

However, the body-part detection algorithm has no way of discriminating the left thigh from the right thigh, or the left shin from the right shin, and if one person is pictured from the back instead of from the front, the left-right limb associations will be erroneous.

Therefore, joining the two thigh regions together, and the two fore-leg regions together as well, improves results in the majority of cases.

Since the results are coherent across the tested datasets, features and NN classifiers, "4 Parts" descriptor extraction is used in the other experiments described in the rest of the chapter.

Another conclusion that is confirmed with these results is that BVT is the most discriminative color feature of the features tested, and that the Hellinger's distance is the most appropriate when using NN classification with the histogram vector features tested. For this reason, NN classification with BVT feature is used as a baseline in many experiments below.

4.2 MULTI-VIEW

The following experiments cover several aspects of the Multi-View classifier. All of the experiments in this Section, unless otherwise noted, are standard RE-ID experiments (no pedestrian detection, single-shot, short term, closed scenario – see Section 1.3 for the definitions), run with the "4 Parts" feature extraction method (see Figure 18d), and using all the test samples as unlabeled data.

4.2.1 *Parameter Selection*

Multi-View optimization has two parameters to be set, g_A that weights the standard RKHS regularization term, and g_I that weights the multi-feature manifold regularization term[2]. Each kernel also has a kernel bandwidth σ parameter to be set.

The $[g_A, g_I]$ parameter space was extensively sampled with the pattern search algorithm [1] in the iLIDS4REID, ViPER and CAVIAR4REID datasets, and the best choice for parameters g_A and g_I, according to the **nAUC** criterion, across all datasets and features, was found to be 0.1 and 10^{-5} respectively. Nevertheless, the parameters g_A and g_I can be optimized once per dataset for increased performance in specific scenarios. In Table 8 the difference in performance from using standard parameters ($g_A = 0.1, g_I = 10^{-5}$) or optimized parameters is displayed. The kernel bandwidth σ is computed on a per-view basis, a "median estimated kernel bandwidth", as described in Section 3.3.1.3.

Except when pointed otherwise the experiments below use $g_A = 0.1$, $g_I = 10^{-5}$, and median estimated kernel bandwidth as parameters.

4.2.2 *Multi-View vs Nearest-Neighbor*

The experiments focused on Multi-View begin by illustrating how allowing Multi-View to train separate classifiers for separate parts of a feature vector (*e.g.*, treating the B, the V and the T parts to the BVT feature vector separately) can outperform applying NN classifier on the same feature vector.

---

2 $$\min_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{i=1}^{l} \|y_i - Cf(x_i)\|^2 + \gamma_A \sum_{i=1}^{l+u} \|f(x_i)\|^2 + \gamma_I \sum_{i=1}^{l+u} \sum_{j,k=1, j<k}^{m} \left\| f^j(x_i^j) - f^k(x_i^k) \right\|^2$$

Dataset	Features	g_A	g_I	**Rank1 (%)**	**Rank5 (%)**	**nAUC**
VIPeR	LBP+MR8+Lab+HSV+BVT	$1e-05$	0.1	19.59	40.76	92.34
	LBP+MR8+Lab+HSV+BVT	$1.9073e-06$	0.10342	19.62	40.89	92.35
iLIDS4REID	LBP+MR8+Lab+HSV+BVT	$1e-05$	0.1	30.76	50.59	86.43
	LBP+MR8+Lab+HSV+BVT	$1.9073e-06$	0.0012213	32.10	52.77	87.86
CAVIAR4REID	LBP+MR8+Lab+HSV+BVT	$1e-05$	0.1	06.40	31.60	72.61
	LBP+MR8+Lab+HSV+BVT	$8.0927e-06$	0.0039307	05.80	31.80	72.64
3DPeS	LBP+MR8+Lab+HSV+BVT	$1e-05$	0.1	21.86	43.32	89.13
	LBP+MR8+Lab+HSV+BVT	0.00059174	0.0037018	23.27	45.93	89.78

Table 8: Results for standard parameters ($g_I = 0.1$, $g_A = 10^{-5}$) and optimized parameters, in four datasets, with five views.

4.2.2.1 *Features used*

The features employed with the NN classifier were:

- Hue-Saturation-Value histogram (HSV)[59];
- Black-Value-Tint histogram (BVT) (Section 4.1.1).

With the Multi-View classifier, those features were decomposed into the following:

- The H part of the HSV feature.
- The S part of the HSV feature.
- The V part of the HSV feature.
- The BV part of the BVT feature.
- The T part of the BVT feature.

All experiments use "4 Parts" descriptor extraction.

4.2.2.2 *Classifiers used*

The NN classifiers employed used the following distances also used in the previous experiment:

- **BhattD** Hellinger's distance,
- **ChiSqD** Chi-Squared distance.

The Multi-View classifier (see Section 3.3.1) was also used, with the features "BV" and "T" as views, and the Bhattacharyya kernel for one experiment, and the features "H", "S" and "V" as views, and the Chi-Squared kernel for another experiment. Both experiments use parameters $g_I = 0.1$, $g_A = 10^{-5}$ and median estimated kernel bandwidth.

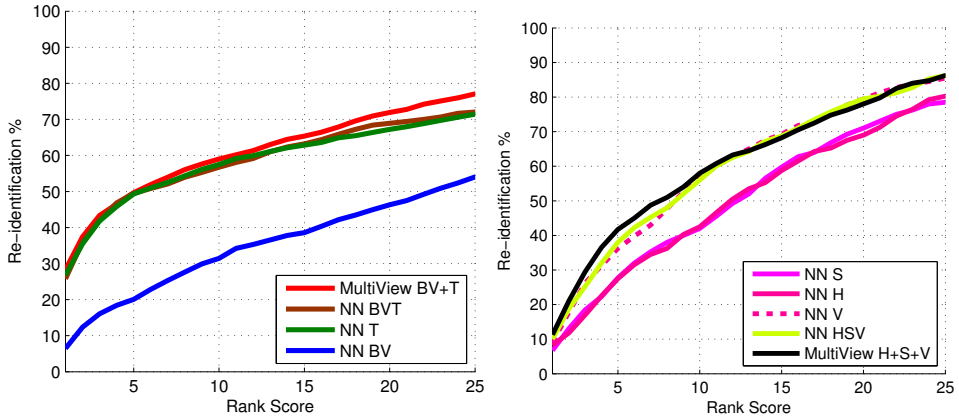
- **BhattK** Bhattacharyya kernel: $K(\mathbf{t}, \mathbf{x}) = exp\left(-\frac{\sqrt{1 - \sum \sqrt{t_i \cdot x_i}}}{\sigma^2}\right)$
- **ChiSqK** Chi-Square kernel: $K(\mathbf{t}, \mathbf{x}) = exp\left(-\frac{\sum \frac{(t_i - x_i)^2}{t_i + x_i}}{\sigma^2}\right)$

4.2.2.3 *Datasets used*

Figure 26: Sample images of a single person from the iLIDS-MA dataset. It contains many images of each pedestrian, from two camera views.

In this experiment the iLIDS4REID and the iLIDS-MA datasets were used (sample images in Figures 24 and 26). iLIDS4REID contains 476 images of 119 pedestrians, captured from up to two different cameras inside an airport. iLIDS-MA contains 3680 images of 40 pedestrians also in the same airport as iLIDS4REID.

For each experiment in each dataset, 10 runs were made, and the results shown are the average of those 10 runs. For each run in the iLIDS4REID and iLIDS-MA dataset, two images were randomly selected from each pedestrian, one to be the probe and one to be in the gallery.



(a) iLIDS4REID dataset, Bhattacharyya distance and kernel, BVT feature vs Multi-View BV+T

(b) iLIDS-MA dataset, Chi-square distance and kernel, HSV feature vs Multi-View H+S+V

Figure 27: Comparison between Multi-View and NN in the iLIDS4REID and the iLIDS-MA datasets. Average results on 10 different data partitions are displayed. Multi-View outperforms NN on average and on each individual partition. Results with the BV, T, H, S and V features are included as a baseline.

Run	Dataset	Features	Equalization Extraction	Descriptor	Classifier	Rank1 (%)	Rank5 (%)	nAUC (%)
1	iLIDS4REID	BVT	×	4 Parts	NN-BhattD	24.2	49.4	85.0
	iLIDS4REID	BV+T	×	4 Parts	MV-BhattK	29.1	49.8	86.6
2	iLIDS4REID	BVT	×	4 Parts	NN-BhattD	25.3	49.0	85.6
	iLIDS4REID	BV+T	×	4 Parts	MV-BhattK	28.8	49.4	86.4
3	iLIDS4REID	BVT	×	4 Parts	NN-BhattD	27.3	48.3	85.3
	iLIDS4REID	BV+T	×	4 Parts	MV-BhattK	29.3	48.5	86.1
4	iLIDS4REID	BVT	×	4 Parts	NN-BhattD	25.1	50.2	85.3
	iLIDS4REID	BV+T	×	4 Parts	MV-BhattK	27.1	50.3	88.0
5	iLIDS4REID	BVT	×	4 Parts	NN-BhattD	27.0	50.0	84.9
	iLIDS4REID	BV+T	×	4 Parts	MV-BhattK	27.7	50.1	87.5
6	iLIDS4REID	BVT	×	4 Parts	NN-BhattD	26.2	49.7	85.8
	iLIDS4REID	BV+T	×	4 Parts	MV-BhattK	27.9	49.8	86.9
7	iLIDS4REID	BVT	×	4 Parts	NN-BhattD	27.4	48.5	85.2
	iLIDS4REID	BV+T	×	4 Parts	MV-BhattK	30.7	48.8	86.1
8	iLIDS4REID	BVT	×	4 Parts	NN-BhattD	24.3	50.7	84.9
	iLIDS4REID	BV+T	×	4 Parts	MV-BhattK	27.8	50.9	86.4
9	iLIDS4REID	BVT	×	4 Parts	NN-BhattD	25.9	50.0	86.0
	iLIDS4REID	BV+T	×	4 Parts	MV-BhattK	28.7	50.1	86.6
10	iLIDS4REID	BVT	×	4 Parts	NN-BhattD	25.0	49.4	85.2
	iLIDS4REID	BV+T	×	4 Parts	MV-BhattK	26.3	49.7	87.1
1	iLIDS-MA	HSV	✓	4 Parts	NN-ChiSqD	9.1	38.2	72.6
	iLIDS-MA	H+S+V	✓	4 Parts	MV-ChiSqK	10.0	41.4	73.3
2	iLIDS-MA	HSV	✓	4 Parts	NN-ChiSqD	8.2	37.0	73.4
	iLIDS-MA	H+S+V	✓	4 Parts	MV-ChiSqK	10.2	43.6	73.8
3	iLIDS-MA	HSV	✓	4 Parts	NN-ChiSqD	9.0	38.9	73.3
	iLIDS-MA	H+S+V	✓	4 Parts	MV-ChiSqK	10.4	41.0	73.4
4	iLIDS-MA	HSV	✓	4 Parts	NN-ChiSqD	10.6	38.3	72.8
	iLIDS-MA	H+S+V	✓	4 Parts	MV-ChiSqK	10.9	41.4	73.5
5	iLIDS-MA	HSV	✓	4 Parts	NN-ChiSqD	10.7	38.1	72.4
	iLIDS-MA	H+S+V	✓	4 Parts	MV-ChiSqK	11.7	41.2	72.8
6	iLIDS-MA	HSV	✓	4 Parts	NN-ChiSqD	10.7	38.5	72.8
	iLIDS-MA	H+S+V	✓	4 Parts	MV-ChiSqK	12.1	40.7	73.6
7	iLIDS-MA	HSV	✓	4 Parts	NN-ChiSqD	10.2	38.3	72.9
	iLIDS-MA	H+S+V	✓	4 Parts	MV-ChiSqK	11.1	41.7	73.7
8	iLIDS-MA	HSV	✓	4 Parts	NN-ChiSqD	10.5	37.1	73.0
	iLIDS-MA	H+S+V	✓	4 Parts	MV-ChiSqK	12.3	41.6	73.5
9	iLIDS-MA	HSV	✓	4 Parts	NN-ChiSqD	9.9	37.8	72.8
	iLIDS-MA	H+S+V	✓	4 Parts	MV-ChiSqK	11.8	43.4	73.6
10	iLIDS-MA	HSV	✓	4 Parts	NN-ChiSqD	11.1	37.9	73.0
	iLIDS-MA	H+S+V	✓	4 Parts	MV-ChiSqK	11.6	41.7	74.1
avg	iLIDS4REID	BV	×	4 Parts	NN-BhattD	06.5	20.1	73.3
	iLIDS4REID	T	×	4 Parts	NN-BhattD	26.8	49.3	84.0
	iLIDS4REID	BVT	×	4 Parts	NN-BhattD	25.8	49.5	85.3
	iLIDS4REID	BV+T	×	4 Parts	MV-BhattK	28.3	49.7	86.8
	iLIDS-MA	H	✓	4 Parts	NN-ChiSqD	08.2	27.5	65.5
	iLIDS-MA	S	✓	4 Parts	NN-ChiSqD	06.8	27.8	65.3
	iLIDS-MA	V	✓	4 Parts	NN-ChiSqD	09.0	36.2	72.2
	iLIDS-MA	HSV	✓	4 Parts	NN-ChiSqD	10.0	38.0	72.9
	iLIDS-MA	H+S+V	✓	4 Parts	MV-ChiSqK	11.2	41.8	73.5

Table 9: NN-BhattD and NN-ChiSqD indicate NN classifier with Hellinger's distance and Chi-Squared distance respectively. MV-BhattK and MV-ChiSqK indicate the Multi-View classifier with Bhattacharyya and Chi-Squared kernel respectively. The performance of BV, T, H, S and V is included to serve as a baseline, to give a sense of the influence of each feature part.

4.2.2.4 *Evaluation Metric*

For these experiments the standard RE-ID metric, the CMC was used. In the figures and tables it is reported the first rank percentage, the fifth rank percentage and the normalized area under the CMC.

4.2.2.5 *Results*

The results in Table 9 and Figure 27 show that utilizing Multi-View to take into account separate views of the features outperforms NN of the simple concatenation of said features. Not only Multi-View outperforms NN on the average of the 10 runs, it also outperforms on each individual run, indicating a robust result.

4.2.3 *Multi-View vs Single view (concatenation of features)*

In the previous Section, the improved performance of Multi-View over NN could be due to the kernel influence. To control for that, in this Section, Multi-View with several features is compared with Multi-View with only one view – the concatenation of the same several features into a single feature vector.

4.2.3.1 *Features used*

The features employed for Multi-View were:

- Local Binary Patterns (LBP) histogram [2].
- Maximum Response Filter Bank (MR8) histogram [75, 109];
- Lightness color-opponent histogram (Lab)[64];
- Hue-Saturation-Value histogram (HSV)[59];
- Black-Value-Tint histogram (BVT) (Section 4.1.1).

For single view, the following were used:

- [**LBP-MR8**] A concatenation of the LBP and MR8 feature vectors.
- [**LBP-MR8-Lab**] A concatenation of the LBP, MR8 and Lab feature vectors.
- [**LBP-MR8-Lab-HSV**] A concatenation of the LBP, MR8, Lab and HSV feature vectors.
- [**LBP-MR8-Lab-HSV-BVT**] A concatenation of the LBP, MR8, Lab, HSV and BVT feature vectors.

All experiments use "4 Parts" descriptor extraction.

4.2.3.2 *Classifiers used*

The Multi-View classifier (see Section 3.3.1) was used, with the Bhattacharyya kernel (see Section 3.3.1.3) with parameters $g_I = 0.1$, $g_A = 10^{-5}$ and median estimated sigmas, for all experiments.

		iLIDS4REID		
	Feature	**R=1**	**R=5**	**nAUC**
SV	LBP	11.60	27.06	74.36
	MR8	12.19	31.85	78.75
	Lab	24.87	46.81	83.42
	HSV	24.37	45.55	83.27
SV	BVT	26.89	47.48	83.96
MV	BV+T	<u>28.31</u>	<u>49.74</u>	<u>86.80</u>
SV	[LBP-MR8]	13.70	34.87	79.85
MV	LBP+MR8	<u>19.92</u>	<u>38.49</u>	<u>81.26</u>
SV	[LBP-MR8-Lab]	22.86	44.96	86.40
MV	LBP+MR8+Lab	<u>26.72</u>	<u>50.08</u>	<u>86.96</u>
SV	[LBP-MR8-Lab-HSV]	25.55	47.82	86.70
MV	LBP+MR8+Lab+HSV	<u>29.50</u>	<u>50.34</u>	<u>86.81</u>
SV	[LBP-MR8-Lab-HSV-BVT]	25.88	48.40	86.13
MV	LBP+MR8+Lab+HSV+BVT	<u>30.76</u>	<u>50.59</u>	<u>86.43</u>

Table 10: Results on the iLIDS4REID dataset, comparing the single vector feature concatenation and the multi-view learning with the respective features. "V" signifies Single View, and "MV" indicates Multi-View. Results show that Multi-View applied to the different features outperforms Multi-View applied to a single vector with the concatenation of features. Best scores are <u>underlined</u>.

4.2.3.3 *Datasets used*

In these experiments the VIPeR and iLIDS4REID datasets were used. For each experiment in each dataset, 10 runs were made, and the results shown are the average of those 10 runs. For each run in the VIPeR dataset, 316 pedestrians were randomly selected, and one image of the pair was taken at random to be the probe and the other to be in the gallery. For each run in the iLIDS4REID dataset, two images were randomly selected from each pedestrians, one to be the probe and one to be in the gallery.

4.2.3.4 *Evaluation Metric*

For these experiments the standard RE-ID metric, the CMC was used. In the tables it is reported the first rank percentage, the fifth rank percentage and the normalized area under the CMC.

4.2.3.5 *Results*

The results expounded in Table 10 and Table 11 show that Multi-View performs better when taking into account the separate features as separate views instead of only using the concatenation of all the feature vectors as a single view.

		VIPeR		
	Feature	**R=1**	**R=5**	**nAUC**
SV	LBP	01.68	7.56	66.93
	MR8	02.02	8.13	73.46
	Lab	11.17	28.51	85.68
	HSV	17.94	38.42	91.61
	BVT	16.71	34.71	88.73
SV	[LBP-MR8]	02.56	07.88	74.28
MV	LBP+MR8	03.39	10.06	76.56
SV	[LBP-MR8-Lab]	06.42	18.92	81.88
MV	LBP+MR8+Lab	10.38	24.87	85.48
SV	[LBP-MR8-Lab-HSV]	14.43	33.83	90.20
MV	LBP+MR8+Lab+HSV	18.01	37.44	91.89
SV	[LBP-MR8-Lab-HSV-BVT]	16.65	36.08	90.88
MV	LBP+MR8+Lab+HSV+BVT	19.59	40.76	92.34

Table 11: Results on the VIPeR dataset, comparing single view and Multi-View. Best scores underlined. LBP, MR8, Lab, HSV and BVT are provided for baseline purposes.

4.2.4 *Multi-View vs NN of Linear Combination of Features*

Here it is explored how MultiView outperforms NN of the linear combination of features. This work uses the NN results already present in [25], comparing the use of the same features in a Multi-View architecture.

4.2.4.1 *Features used*

The features employed were:

- Maximally Stable Color Regions (MSCR) histogram [51].
- Black-Value-Tint histogram (BVT) (Section 4.1.1).

All experiments use "4 Parts" descriptor extraction.

4.2.4.2 *Classifiers used*

The NN classifier was employed with the following distance:

- **Bhatt** Hellinger's distance: $D(x,t) = \sqrt{1 - \sum_{i=1}^{E} \sqrt{x_i \cdot t_i}}$.

The features BVT with MSCR are combined linearly such as in [25] Cheng *et al.* First, the features are extracted from all images, then an all-to-all distance matrix is computed for BVT (D_{BVT}) and another for MSCR (D_{MSCR}). Then these matrices are linearly combined as follows: $D = 0.7D_{BVT} + 0.3D_{MSCR}$ (the weights 0.7 and 0.3 were determined by exhaustive search). Then NN classification is performed on top of the resulting distance matrix D, where

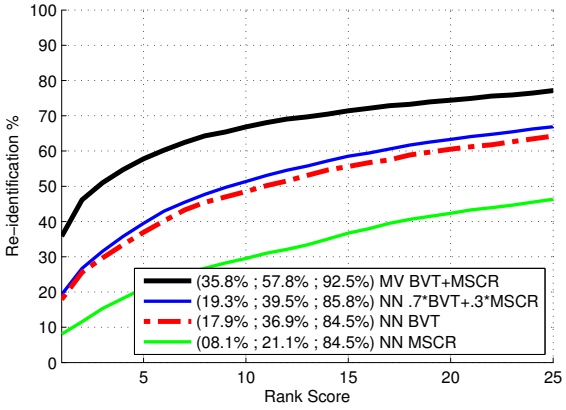


Figure 28: Comparison between Multi-View and NN of linear combination of features. Results illustrated in the VIPeR dataset, with Bhattacharyya distance for NN and kernel for Multi-View. BVT and MSCR are included for baseline purposes.

the minimum distance from each probe to all gallery images if found, to determine the nearest-neighbor match for each probe image.

The Multi-View classifier (see Section 3.3.1) was also used, combining the same BVT and MSCR features. The Bhattacharyya kernel (see Section 3.3.1.3) was used with parameters $g_I = 0.1$, $g_A = 10^{-5}$ and median estimated sigmas, for all experiments.

$$\text{BHATTACHARYYA KERNEL: } K(\mathbf{t}, \mathbf{x}) = exp\left(-\frac{\sqrt{1 - \sum \sqrt{t_i \cdot x_i}}}{\sigma^2}\right)$$

4.2.4.3 *Datasets used*

In these experiments the VIPeR dataset was used. For each experiment 10 runs were made, and the results shown are the average of those 10 runs. For each run in the VIPeR dataset, 316 pedestrians were randomly selected, and one image of the pair was taken at random to be the probe and the other to be in the gallery.

4.2.4.4 *Evaluation Metric*

For these experiments the standard RE-ID metric, the CMC was used. In the figure it is reported the first rank percentage, the fifth rank percentage and the normalized area under the CMC.

4.2.4.5 *Results*

Figure 28 clearly shows that Multi-View is better able to integrate two features than NN of the linear combination of them. It is not shown, but not only does Multi-View outperform NN in the average of the 10 partition runs, but also on each individual run.

4.2.4.6 *Discussion*

This experiment suggests that MultiView effectively trains better classifiers for each feature than nearest neighbor of the linear combination of each feature.

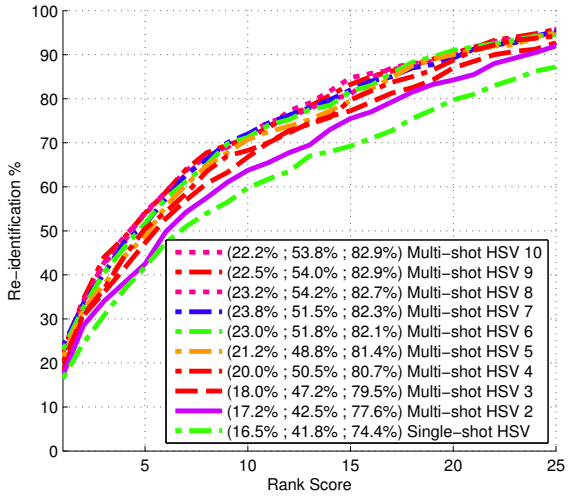


Figure 29: Using shots in a Multi-Shot scenario as views of Multi-View. Each line represents a Multi-Shot with N shots and views, where each view is the feature histogram of the person image in one shot. The rank 1, rank 5 and **nAUC** percentages are also displayed. This test was performed in the iLIDS-MA dataset, with HSV feature, and the Bhattacharyya kernel.

		R=1	R=5	nAUC
MFL	Multi-shot 10	22.25	53.75	82.94
	Multi-shot 9	22.50	54.00	82.87
	Multi-shot 8	23.25	54.25	82.66
	Multi-shot 7	23.75	51.50	82.33
	Multi-shot 6	23.00	51.75	82.09
	Multi-shot 5	21.25	48.75	81.38
	Multi-shot 4	20.00	50.50	80.66
	Multi-shot 3	18.00	47.25	79.46
	Multi-shot 2	17.25	42.50	77.56
SFL	Single-shot	16.50	41.75	74.39

Table 12: Results on the iLIDS-MA dataset, comparing the performance level between different number of "shots" as views in a multi-shot scenario. "SLF" signifies Single Feature Learning, and "MFL" indicates Multi Feature Learning.

4.2.5 *Views as any facet of a target*

Here it is explored the hypothesis that multi-shot yields a superior re-identification performance than single-shot, when using Multi-View classification. In this case, each view represents the features extracted from a different image of a

given pedestrian in a multi-shot scenario (where there are several images per exemplar).

These are standard RE-ID experiments (no pedestrian detection, short term, closed scenario – see Section 1.3), ran in multi-shot scenarios.

4.2.5.1 *Features used*

The feature employed was:

- Hue-Saturation-Value histogram (HSV)[59];

All experiments use "4 Parts" descriptor extraction, and each feature when applied to a region of the image generate a histogram of constant bin size for all experiments (illustrated in Figure 22).

4.2.5.2 *Classifiers used*

The Multi-View classifier (see Section 3.3.1) was used, with the Bhattacharyya kernel (see Section 3.3.1.3) with parameters $g_I = 0.1$, $g_A = 10^{-5}$ and median estimated kernel bandwidth, for all experiments.

4.2.5.3 *Datasets used*

In this experiment the iLIDS-MA datasets was used (sample images in Figure 26). iLIDS-MA contains 3680 images of 40 pedestrians also in the same airport as iLIDS4REID. It was at the time of the experiment, one of the few datasets with more than 20 samples per pedestrian, such as to allow multi-shot with 10 samples per exemplar.

There were 10 experiments, each with an increasing number of samples per pedestrian, from 1 sample (single-shot case) up to 10 samples (multi-shot cases). For each experiment, 10 runs were made, and the results shown are the average of those 10 runs. For each run, where N was the number of samples per pedestrian to be used, $2 \times N$ images were randomly selected from each pedestrians, N to be the probes and N to be part of the gallery.

4.2.5.4 *Evaluation Metric*

For these experiments the standard RE-ID metric, the CMC was used. In the figure and table it is reported the first rank percentage, the fifth rank percentage and the normalized area under the CMC.

4.2.5.5 *Results*

Each experiment depicted in Figure 29 and Table 12 represents a different number of views, each view an additional image for each pedestrian sample.

In "Multi-shot 2", each pedestrian has 2 images per sample (two images for the gallery sample and 2 for the probe sample). In "Multi-shot 3", each person has 3 images per sample, and so on.

The experiment depicted in Figure 29 and Table 12 consistently obtains better results when using multi-shot than single-shot, and using more samples

consistently improves the **nAUC** performance, even if only slightly. As for rank 1, using more samples improves results up to a point, after which it stagnates.

4.2.5.6 *Discussion*

The experiment is one more indicator (of many *e.g.*, [25, 42]) that multi-shot yields a superior re-identification performance than single-shot, as is intuitive since there are more images and therefore more information to be exploited. It also adds to the many experiments where Multi-View successfully integrates information from various sources. Concurrently it illustrates how with Multi-View classification, each view need not be a feature, but may be any facet of the given data samples – in this case, each view represents the information extracted from an image of the given pedestrians in the multi-shot scenarios.

4.2.6 *Comparison with other Re-Identification algorithms*

Here experiments are carried out to compare the Multi-View classifiers with other state-of-the-art techniques.

4.2.6.1 *Features used*

The features employed were:

BVT Black-Value-Tint histogram (BVT) (Section 4.1.1).

HSV Hue-Saturation-Value histogram (HSV)[59];

LAB Lightness color-opponent histogram (Lab)[64];

MR8 Maximum Response Filter Bank (MR8) histogram [75, 109];

LBP Local Binary Patterns (LBP) [2].

All experiments use "6 Parts" descriptor extraction, and each feature when applied to a region of the image generate a histogram of constant bin size for all experiments (illustrated in Figure 22).

4.2.6.2 *Classifiers used*

The Multi-View classifier (see Section 3.3.1) was used for the single-frame classifier, with the Bhattacharyya kernel (see Section 3.3.1.3) for all experiments.

$$\text{BHATTACHARYYA KERNEL: } K(\mathbf{t},\mathbf{x}) = exp\left(-\frac{\sqrt{1-\sum\sqrt{t_i \cdot x_i}}}{\sigma^2}\right)$$

In Table 13, for the **MFL** experiments, the regularization parameters are set to $\gamma_I = 10^{-5}$ and $\gamma_A = 0.1$, the kernel parameters (σ^i) were estimated as noted in Section 3.3.1.3. For the **MFL opt.** experiments the regularization parameters along with the kernel parameters were optimized using the pattern search algorithm [1].

4.2.6.3 *Datasets used*

Figure 30: Sample images from the CAVIAR4REID dataset. It contains a ten images of each pedestrian from each camera, from up to two cameras in a shopping center with very low resolution.

In these experiments the iLIDS4REID, the VIPeR, and the CAVIAR4REID datasets were used (sample images in Figures 24, 23 and 30). CAVIAR4REID contains 1220 images captures in a shopping center environment. It containst 10 images per camera, of two cameras, per individual of 50 persons, and 10 more images per person of 22 pedestrians that only appear in one of the cameras.

Each dataset was randomly split 10 times in gallery and probe sets, and the results shows the average of the results over the different trials. In these experiments, the probe set is considered as unlabeled data.

4.2.6.4 *Evaluation Metric*

For these experiments the CMC was used as a metric. In the figures it is reported the first rank percentage, the fifth rank percentage, the tenth rank percentage, the twentieth rank percentage and the normalized area under the CMC.

4.2.6.5 *Results*

The results presented in Table 13 show that Multi-View compares favourably with several state-of-the-art algorithms. **MFL opt.** outperforms all the methods in terms of nAUC and almost all the reported points of the CMC. PS is slightly better than **MFL opt.** in a few points: $r = \{10, 20\}$ in VIPeR and $r = 1$ in CAVIAR4REID. In general, **MFL opt.** outperforms PS when considering the overall statistics on the CMC, such as the nAUC.

4.2.7 *Comparison with other Semi-Supervised Algorithm*

Re-Identification is a field where there are often more unlabeled samples than labeled ones. This suggests the use of semi-supervised algorithms to exploit all this unlabeled data for increased performance.

iLIDS

	r = 1	r = 5	r = 10	r = 20	nAUC
SDALF [42]	28.49	48.21	57.28	68.26	84.99
PS [25]	27.39	**52.27**	*60.92*	*71.85*	*87.08*
[92]	25.97	43.27	55.97	67.31	83.14
[127]	24.00	43.50	54.00	66.00	–
MFL	*30.76*	50.59	58.74	70.42	86.44
MFL opt.	**31.51**	*51.18*	**62.43**	**74.79**	**88.40**

VIPeR

	r = 1	r = 5	r = 10	r = 20	nAUC
SDALF [42]	19.87	38.89	49.36	65.72	92.24
PS [25]	*21.17*	*42.66*	**56.90**	**72.82**	*93.51*
RDC [128]	15.66	38.42	53.86	70.09	–
[82]+RankSVM [104]	15.73	37.66	51.17	66.27	–
[82]+RDC [128]	16.14	37.72	50.98	65.95	–
MFL	19.59	40.76	52.21	66.11	92.34
MFL opt.	**22.53**	**44.40**	*55.92*	*70.70*	**93.75**

CAVIAR4REID

	r = 1	r = 5	r = 10	r = 20	nAUC
SDALF [42]	6.80	25.00	44.40	65.80	68.65
PS [25]	**8.60**	30.80	47.80	*71.60*	72.38
MFL	6.40	*31.60*	*48.20*	70.60	*72.61*
MFL opt.	*8.20*	**35.20**	**53.20**	**74.00**	**74.39**

Table 13: Results on iLIDS (top), VIPeR (middle) and CAVIAR4REID datasets (bottom), comparing the Multi-View classifier with state of the art classifiers. Best scores in bold, second best scores in italic.

The work in [86] used unlabeled images from a camera pair to to exploit the geometry of the marginal distribution for obtaining robust sparse representation. Another approach, that of [82], used unlabeled images to discover clusters where some feature is more informative than all others, to then exploit this information in the test phase. Additionally, [83] uses unsupervised clustering forests to propagate human input to the rest of the unlabeled samples. Recently, [89] explored the issue of very few labeled samples during the training stage.

Here MultiView is compared with another semi-supervised algorithm in the state of the art, the work of [24].

4.2.7.1 *Features used*

The features employed were:

- Hue-Saturation-Value histogram (HSV)[59];

All experiments use "4 Parts" descriptor extraction, and each feature when applied to a region of the image generate a histogram of constant bin size for all experiments (illustrated in Figure 22).

4.2.7.2 *Classifiers used*

The Multi-View classifier (see Section 3.3.1) was used, with the Bhattacharyya kernel (see Section 3.3.1.3) for all experiments.

Cabral's *et al.* work [24] was chosen for this comparison because it is a recent algorithm in the state of the art with competitive results, and more importantly, the author provided the code on request. It provides matrix completion. Given specially constructed matrices with labels alongside features set in columns for training, and similarly for the test samples, columns of features but with unknowns in the position of the labels, the algorithm fills in the label slots with the person classifications. The MC-Simplex algorithm described in [23] was used. The parameters γ and λ were fine tuned in the range $\gamma \in [1, 30], \lambda \in \left[10^{-4}, 10^{2}\right]$, and the μ threshold was set to 10^{-9} as described in [23].

4.2.7.3 *Datasets used*

In this experiment the iLIDS-MA datasets was used (sample images in Figure 26). iLIDS-MA contains 3680 images of 40 pedestrians also in the same airport as iLIDS4REID. It was at the time of the experiment, one of the few datasets with more than 20 samples per pedestrian, such as to allow multi-shot with 10 samples per exemplar.

There was one experiment, with 10 samples (a multi-shot case). For each experiment, 10 runs were made, and the results shown are the average of those 10 runs. For each run, 20 images were randomly selected from each pedestrians, 10 to be the probes and 10 to be part of the gallery.

4.2.7.4 *Evaluation Metric*

For these experiments the standard RE-ID metric, the CMC was used. In the figures it is reported the first rank percentage, the fifth rank percentage and the normalized area under the Cumulative Matching Characteristic curve.

4.2.7.5 *Results*

The experiment depicted in Figure 31 illustrates how there are successful semi-supervised algorithms in other fields (*i.e.*, Cabral's *et al.* work in Image Categorization [24]) that Multi-View (and NN for that matter) outperform in the re-identification field.

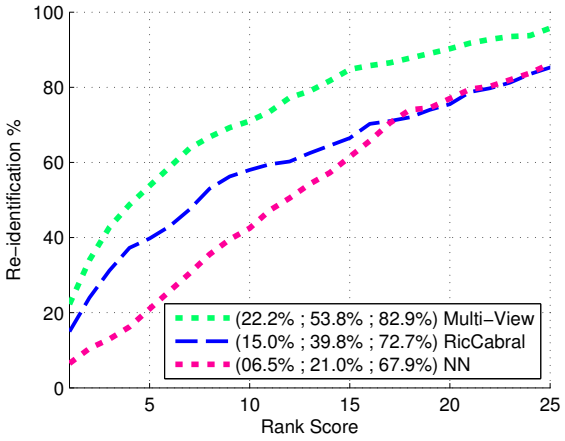


Figure 31: Comparison with another semi-supervised algorithm (Ricardo Cabral's Matrix Completion algorithm [24]). Tests run in the iLIDS-MA dataset, with the HSV feature, in a multi-shot scenario with 10 shots. NN is included as a baseline.

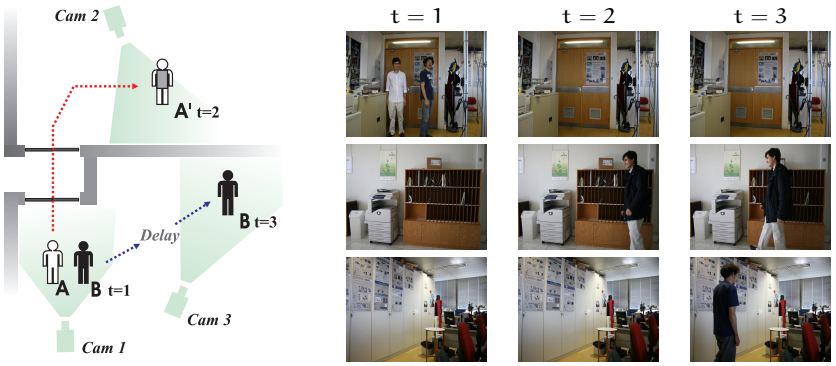
4.2.8 *Discussion on the Theoretical Differences of Multi-View and State-of-the-Art Algorithms*

Multi-View learns weights for each sample and feature. The proposed Multi-View formulation allows for each feature to have a different distance metric in its kernel, but as is, these distance metrics must be set before optimization. On the other hand, [101] optimizes over different distance metrics to find which best fit each sample. In principle, Multi-View can be readily extended to also learn weights for different distance metrics for each sample. The classifier definition could be extended to not only do a weighted sum of a kernel response for each training sample against the test sample, but to also include more weighted sums of other kernels for each training sample.

[84] however, by adapting the feature weights *on-the-fly* for each probe sample is on a qualitatively different level. Multi-View is limited to a global outlook, setting its weights during training instead of on-the-fly.

4.3 MULTIPLE HYPOTHESES TRACKING

The problem of tracking people across a camera network is often addressed with re-identification only, *i.e.*, matching through the similarity between each detection and each existing target. Few works actually use inter-camera tracking mechanisms. One such work, also in the context of tracking in camera networks, is the work of Javed et al. [66], that uses a Maximum a Posteriori (MAP) approach, similar to a global nearest neighbor's association [15]. In the conducted experiments here, the MHT algorithm in its standard implementation, is compared with an MHT implementation with only one leaf, which is equivalent to the MAP approach (as defined by [15]). MAP considers all detections and existing targets at each scan and chooses the best assignment.



(a) (b)

Ground truth localiz.	t = 1	t = 2	t = 3
Zone 1	A, B		
Zone 2		A	A
Zone 3			B

(c)

MAP localiz.	t = 1	t = 2	t = 3
Zone 1	A, B		
Zone 2		B	B
Zone 3			A

(d)

MHT localiz.	t = 1	t = 2	t = 3
Zone 1	A, B		
Zone 2		B	A
Zone 3			B

(e)

Figure 32: Two people tracking with three non-overlapping cameras. Persons A and B start in the field of view of Cam1, and then both move out. A puts on a jacket and enters the field of view of Cam2. After some delay, B enters in the field of view of Cam3 (a). Top, middle and bottom rows show images acquired by Cam1, Cam2 and Cam3, respectively (b). Ground truth localization of people (c), estimated localization using MAP (d) and using the proposed MHT (e).

However, it does not account for the possibility that the assignment may be erroneous [15].

4.3.1 *Illustrative Example: Changing target*

In this experiment, the tracking area includes three zones, z_1, z_2, and z_3, corresponding to the fields of view of three cameras, Cam1, Cam2, and Cam3, respectively (see Figure 32 (a)). Figure 32 (b) shows just three images for each camera, but in fact there are many more intermediate images. The time stamps, $t = 1$, $t = 2$, and $t = 3$, indicate relevant events, namely beginning of experiment and appearance of novel objects in the cameras of the network. The video frames captured by the three cameras are processed in order to detect foreground objects and detected objects are characterized through HSV color histograms, with 10 bins for each channel, as shown in Figure 22, with

the 2 Part descriptor extraction that divides a person body in two by the waist (Section 3.2).

In the beginning of the experiment two persons, A and B, are visible in z_1, and both walk away, leaving the field of view of Cam1. Then, A appears in Cam2 and, shortly after, B appears in Cam3. The person A is initially wearing white clothes, while B is wearing dark clothes (see the top-left image in Figure 32 (b)). When A reappears in Cam2 he is wearing a dark jacket, changing his color histogram significantly. With the jacket he becomes more similar to B, as seen in Cam1, than with himself.

Tables (c), (d) and (e), in Figure 32, show the ground truth, the tracker predictions of MAP and MHT respectively. At $t = 2$, both MAP and MHT algorithms make an incorrect association, placing B in z_2. However, at $t = 3$, *i.e.*, when B later appears in Cam3 (z_3), MHT is able to correct the prediction, and thus concludes that A went to z_2 and B went to z_3, while MAP maintains the incorrect association.

The rationale behind the correction of the MHT prediction is as follows. The color description of the persons is not expected to change, therefore, hypotheses in which the histograms change receive a probability penalization. This penalization occurs via the $P_{h^{z},h^{\tau}}$ term of $P_{Z_t^k,{}^{\iota}T_i^k}$. Assume now a simplification with grey scale histograms having only one bin, which will be used to give the reader the intuition of what is being calculated by the MHT algorithm and why it is able to correct its previous decision. In all experiments, the targets' gates are of size 1, thus the tracker assigns a probability of zero to the possibility of a target crossing a zone undetected.

For simplicity let's assume a histogram feature of a single bin. In Cam1, assume that A has a "histogram" with a value of 0 in its single bin, and B a "histogram" with a value of 1. When A appears in Cam2 he has a histogram of 0.7. In the hypothesis according to which A is in z_2, the total change in histograms is of 0.7, but in the hypothesis which places B in z_2, the total change in histograms is only 0.3. Thus, at this point, B would always be placed in z_2 by any algorithm. When B appears in Cam3, his histogram in that detection is still 1. Because the target's gate is 1, when one of the persons is in z_2, the algorithm will place the other person in z_3, that is, in the hypothesis where A is in z_2, B will be placed in z_3 (z_3 is not in A's gate), in the hypothesis where B is in z_2, A will be placed in z_3.

The hypothesis which placed B in z_2 has a total change in histograms of $0.3 + 1 = 1.3$, but the hypothesis which correctly placed A in z_2 has a total change in histograms of only 0.7. Because greater change in histograms directly translates into lower probability of an hypothesis, the hypothesis which placed A in z_2 will now be selected as the best hypothesis, because it has a total change in the histograms of only 0.7, versus 1.3 in the other hypothesis.

With MAP, person A would be incorrectly labeled as B in z_2, and when B really appeared in z_3, the best assignment would be to incorrectly place A in z_3. Furthermore, if no tracking algorithm was used, then B could possibly be assigned to the detection in z_2, and *also* to the detection in z_3.

4.3.2 *Simulation*

A large tracking area is simulated, consisting of 57 zones, each zone containing a camera (depicted in Figure 33 (a)). During the simulation, 40 targets move in the tracking area. Each target initially chooses a random zone and walks there by the shortest path. Upon arrival, he repeats the same behavior, indefinitely. Two sources of uncertainty are considered. One source of uncertainty models camera noise, illumination changes, person pose and other changes alike, as additive Gaussian noise in the targets' histogram. The other source of uncertainty models target detector reliability issues by deleting from the simulation detections at a certain mis-detection rate.

Several simulations were run, with varying values of histogram noise and mis-detection rates. Each simulation is comprised of 5000 scans, with 1 second per scan, and the simulated people take 3 to 6 scans to move between areas. The history of the tracks produced by each tracker is analyzed and the average number of incorrect assignments per scan during the simulation is used to measure the tracker's performance. If a target ιT^{k-1} was assigned an identifier i in scan $k-1$ and an identifier $j \neq i$ in scan k by the tracking algorithm, then an assignment error occurred.

In Figure 33 (e), the performance of MHT and MAP are presented, with varying levels of noise added to the target's histograms, for a percentage of misdetections of 15%. In figure 33 (f) the percentage of misdetections is varied, for an added noise in the target's histograms of 0.8.

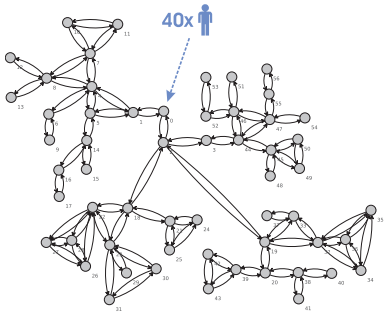
Comparing the MHT's performance with the MAP performance, MAP makes the best assignment between measurements and targets at each scan but, as it does not maintain multiple hypotheses on possible states of the world, it cannot recover from past mistakes as well as MHT does. Therefore, MHT consistently obtains better results than the MAP approach.

4.4 INTEGRATING PEDESTRIAN DETECTION AND RE-ID

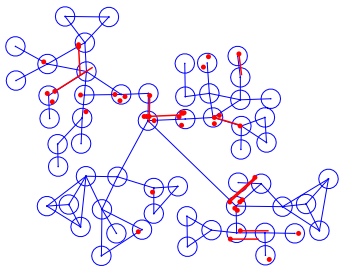
This section presents an extensive evaluation of the proposed system for integrating automatic pedestrian detection in RE-ID. The experimental evaluation will give emphasis to the novelties presented in the framework, namely: *(i)* the influence of the occlusion filter and the false positive class presented in Section 3.1.2 and Section 3.1.3; *(ii)* the performance of our window-based RE-ID classifier for all the combination of parameters (r,d,*w*) comparing against the respective single-frame classifier ($\mathcal{T}$=(r, 1, 1)).

4.4.1 *Evaluation*

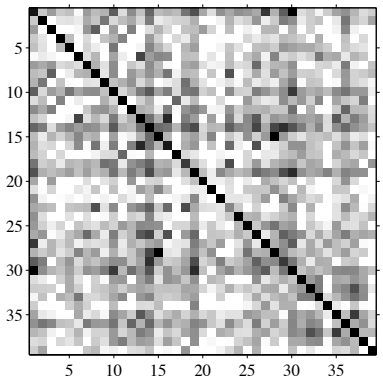
The integration of PD and RE-ID generates more types of errors than the regular RE-ID experiments, which necessitates additional metrics beyond the CMC for a full evaluation. Here, the re-identification evaluation method utilized in this section is presented.



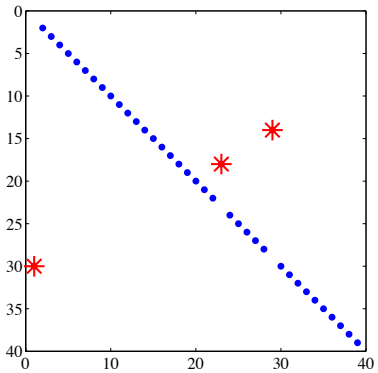
(a) Setup formed by 57 zones and 40 targets



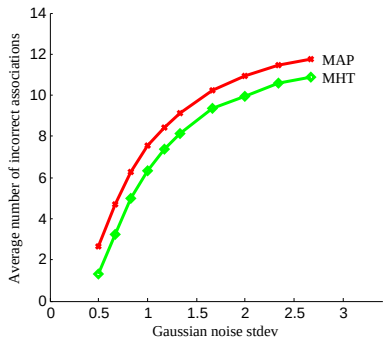
(b) Targets motion from $t = 819$ to $t = 820$



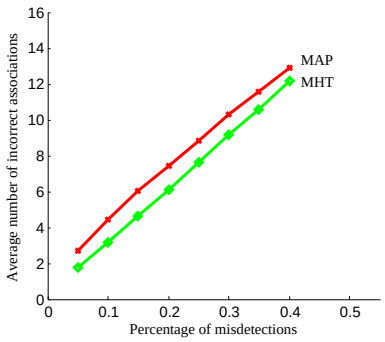
(c) Target distances from $t = 819$ to $t = 820$



(d) Minimal distances from $t = 819$ to $t = 820$



(e) Association errors vs noise added to target histograms



(f) Association errors vs % of misdetections

Figure 33: Simulated experiment involving the tracking of 40 targets in a 57 zones setup (a). All the targets can move to an adjacent node at each time step (b). Distances $1 - B(h^Z, h^T)$ (Eq. 4.1.2) and the best matchings among all targets are shown in (c) and (d) for the same time step indicated in (b). Correct and incorrect histogram matchings are marked with blue dots and red stars, respectively. Assessment of target and measurement associations, using MAP and MHT considering noise in the observed histograms (e) or a varying percentage of misdetections (f).

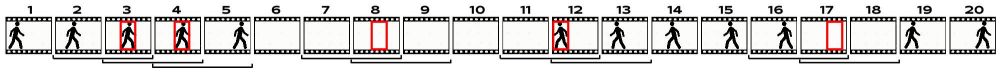


Figure 34: Illustration of $\mathbf{Prec_f}$ and $\mathbf{Rec_f}$ metrics calculation. The pedestrian of interest appears in 12 frames of this video. Each red bounding box indicates a detection and re-identification of rank 1 of that pedestrian. In this example, I set the minimum number of re-identifications to d=1 and the window size to w=2. Given the detections and these parameters, the black brackets below frames {2, 3, 4, 5, 7, 8, 9, 11, 12, 13, 16, 17, 18} indicate the 13 frames that are shown as output of the system. From these 13 frames shown, 7 of them truly contain the pedestrian of interest, therefore precision in frames ($\mathbf{Prec_f}$) is 7/13. From the 12 frames in which the pedestrian appears in the video, only 7 are shown, thus recall in frames ($\mathbf{Rec_f}$) is 7/12. Note how although the detection and re-identification in frame 17 is *erroneous* (a false-positive of the detector, and a lucky mis-classification of the re-identifier), the corresponding video-clip shown *does indeed* contain the pedestrian of interest, and thus is a positive video-clip and thus contributes positively for the recall.

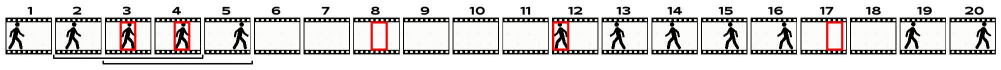


Figure 35: Visualization of the impact that different parameters have on the $\mathbf{Prec_f}$ and $\mathbf{Rec_f}$ metrics, by comparison with the previous figure. In this example, I set the minimum number of re-identifications d=2 and the window size w=3. Given the detections and these parameters, the black brackets below frames {2, 3, 4, 5} indicate the 4 frames that are shown as output of the system. From these 4 frames shown, all 4 of them truly contain the pedestrian of interest, therefore precision in frames ($\mathbf{Prec_f}$) is 1. From the 12 frames in which the pedestrian appears in the video, only 4 are shown, thus recall in frames ($\mathbf{Rec_f}$) is 4/12.

The standard metric for Re-Identification (RE-ID) evaluation is the Cumulative Matching Characteristic curve (CMC), that shows how often, on average, the correct person ID is included in the best r matches against the gallery, for each probe image. If $\mathrm{ord}(i)$ is defined as the number of correct re-identifications at index i in the ordered list of all matches for a probe sample against all classes in the gallery, then CMC is defined as:

$$\mathbf{CMC}(r) = \sum_{i=1}^{r} \frac{\#\mathrm{ord}(i)}{tp}, \quad r \in [1, \ldots, \#\text{ of classes}] \tag{11}$$

where tp is the true positives of the detector, and thus the total number of probes.

This means that when there are False Positive (FP) probes, without a FP class, each FP contributes to the denominator of Equation 11 (see Equation 12) in the CMC calculation, reducing every value of the CMC by the fraction of the amount of FPs relative to the total of probes.

$$\mathbf{CMC}'(r) = \sum_{i=1}^{r} \frac{\#\mathrm{ord}(i)}{tp + FP} = \mathbf{CMC}(r)\frac{tp}{tp + FP} < \mathbf{CMC}(r) \tag{12}$$

When there are MDs, if on average the samples missed are distributed proportionally to $\mathrm{ord}(i)$, then the CMC does not change. This means that the CMC does not penalize the Missed Detections (MDs) introduced by the Pedestrian Detection (PD) algorithm. If there are MDs, there are less probes to be classified (numerator) and the CMC values are divided by a smaller number of probes (numerator).

$$\begin{aligned}\mathbf{CMC}''(r) &= \sum_{i=1}^{r} \frac{\#\mathrm{ord}(i) - \#\mathrm{ord}(i)\frac{MDs}{tp}}{tp - MDs} \\ &= \sum_{i=1}^{r} \frac{\#\mathrm{ord}(i)(1 - \frac{MDs}{tp})}{tp - tp\frac{MDs}{tp}} \\ &= \sum_{i=1}^{r} \frac{\#\mathrm{ord}(i)(1 - \frac{MDs}{tp})}{tp(1 - \frac{MDs}{tp})} = \mathbf{CMC}(r)\end{aligned}$$

Therefore, to take into account both the MDs and FPs introduced by the pedestrian detector, other metrics should be used to complement the performance evaluation of a automatic Re-Identification (RE-ID) system.

In other fields such as object detection and tracking, precision and recall metrics are used to evaluate the algorithms[3]. Here we take inspiration from such examples and adapt precision and recall metrics to evaluate not only the detection part but also the integrated RE-ID and PD system.

Let a certain query for a person i, $i \in 1 \cdots P$, result in N^i presented videos v_n^i, $n \in 1 \cdots N^i$. Let $t(v_n^i)$ be the total number of frames in the video, and $p(i, v_n^i)$ be the number of frames actually containing person i. Finally, let $gt(i)$ be the correct number of frames where pedestrian i appears in the whole sequence.

- **Precision in frames ($\mathbf{Prec_f}$):** Number of frames shown that do contain the pedestrian of interest over the total number of frames shown.

$$\mathbf{Prec_f} = \frac{1}{P}\sum_{i=1}^{P} \frac{\sum_{n=1}^{N^i} p(i, v_n^i)}{\sum_{n=1}^{N^i} t(v_n^i)}$$

- **Recall in frames ($\mathbf{Rec_f}$):** Number of frames shown that do contain the pedestrian of interest over the correct total number of frames in which the pedestrian appears.

$$\mathbf{Rec_f} = \frac{1}{P}\sum_{i=1}^{P} \frac{\sum_{n=1}^{N^i} p(i, v_n^i)}{gt(i)}$$

See Figure 34 and Figure 35 for two illustrative examples and note the variation of the performance metrics in the same video for different d and w.

3 Such as in the iLIDS dataset's user guide: http://www.siaonline.org/SiteAssets/Standards/PerimeterSecurity/iLidsUserGuide.pdf

To summarize, there are several metrics, and they may be combined in any number of ways to provide a final performance measure. Recall penalizes MDs and thus if the application absolutely requires to have the minimum possible of MDs (*i.e.*, detecting strangers in a high-security research facility), recall should be given higher weight. Precision penalizes FPs and thus if the application favors not providing too much wrong output (*i.e.*, video surveillance in a shopping mall, where the confidence of the human operators in the system is considered more important than the security level) then precision should have more weight. Precision in frames also penalizes positive video-clips that only have a few frames containing the pedestrian of interest, so it also accounts for the attentional load put on the user.

One of the often utilized combined metric is the F-score defined below:

$$\textbf{F-score} = 2.\frac{\textbf{Prec}_\textbf{f}.\textbf{Rec}_\textbf{f}}{\textbf{Prec}_\textbf{f} + \textbf{Rec}_\textbf{f}} \tag{13}$$

The F-score is the harmonic mean of $\textbf{Prec}_\textbf{f}$ and $\textbf{Rec}_\textbf{f}$ and is a classical way to combine precision and recall.

4.4.2 *Features used*

The features employed were:

BVT Black-Value-Tint histogram (BVT) (Section 4.1.1).

HSV Hue-Saturation-Value histogram (HSV)[59];

LAB Lightness color-opponent histogram (Lab)[64];

MR8 Maximum Response Filter Bank (MR8) histogram [75, 109];

LBP Local Binary Patterns (LBP) [2].

All experiments use "4 Parts" descriptor extraction, and each feature when applied to a region of the image generate a histogram of constant bin size for all experiments (illustrated in Figure 22).

4.4.3 *Datasets used*

In this final experiments the HDA dataset [116][4] was used (sample images in Figure 36). This is one of the most challenging datasets available. It is the only one up to this point that provides high-definition images. It provides the largest amount of camera views. And provides plenty of challenging examples of varying illumination, occlusion and even changing clothes. It contains over 64'000 images of 85 pedestrians, viewed from up to 13 camera views in an office space scenario. Each video sequence acquired from each camera corresponds to 30 minutes of video during rush hour in our laboratory facilities.

A closed-space assumption is considered for the experimental setup (see Section 1.3) and there is gallery samples for all the pedestrians in the video.

4 http://vislab.isr.ist.utl.pt/hda-dataset

Figure 36: Sample images from the HDA dataset. It contains many images of very different scales, from VGA up to 4MPixel cameras. From up to thirteen different camera views in a office space environment. It includes the notable challenge of changing apparel.

A set of images is collected before-hand and stored in a gallery associated to their identities. Two disjoint sets are used for gallery and probes. More specifically, the best[5] images of 12 out of the 13 cameras sequences were selected for the gallery, and the left-out sequence is used as a probe set. The gallery is built by hand-picking one manually cropped bounding box image for each pedestrian in the sequences that they appear, leading to a total of 230 cropped images for 76 pedestrians (roughly three images per pedestrian). Having, on average, three high quality images for each individual is realistic for a real-life controlled entry point – a few cameras can be set to point at the entry point to capture high-quality images from distinct points of view.

The False Positive (FP) class (Section 3.1.3) is built with the detections from the gallery sequences that have no overlap with any Ground Truth (GT) Bounding Box (BB), for a total of 3972 detections in the FP class. In a realistic case, the system could be set to work automatically by acquiring images early in

5 Best is here defined as images of pedestrians with full visibility and closest to the camera.

the morning, when the building is known to be empty, collect all detections of pedestrians, which will all be FPs, and construct the FP class.

The probe image sequence contains 1182 GT BBs, centered on 20 different people. Such people are fully visible in 416 occurrences and appear occluded in some degree by other BBs, or truncated by the image border, in 766 occasions. Since three pedestrians who appear in the probe set are not present in the gallery set, we remove their corresponding 85 appearances from the probe set (leaving 1097 appearances). The remaining 17 individuals cross the field of view of the probe camera 54 distinct times, therefore there are 54 GT video-clips[6]. Figure 37 displays in blue the appearances throughout the video of each of the 17 pedestrians.

4.4.3.1 *Pedestrian Detection*

In this work we used our implementation [114] of Dollár's Fastest Pedestrian Detector in the West [38] (FPDW). Being FPDW a monolithic detector, it is constrained to generate detections which lie completely inside the image boundary. This naturally generates a detection set without persons truncated by the image boundary, facilitating the RE-ID. This module outputs 1182 detections[7] on the probe camera sequence. The initial detections are filtered based on their size, removing the ones whose height is unreasonable given the geometric constraints of the scene (under 68 pixels). This rejects 159 detections and allows 1023 of them pass. The three pedestrians who appear in the probe set and are not present in the gallery set generate 59 detections which we remove from the detections' pool, since they violate the closed-space assumption. This leads to the 964 elements that form the base set of detections, 155 of which are FPs. Figure 37 displays in green the detections throughout the video of each of the 17 pedestrians.

4.4.4 *Classifiers used*

The Multi-View classifier (see Section 3.3.1) was used for the single-frame classifier, with the Bhattacharyya kernel (see Section 3.3.1.3) for all experiments.

$$\text{BHATTACHARYYA KERNEL: } K(\mathbf{t}, \mathbf{x}) = exp\left(-\frac{\sqrt{1 - \sum \sqrt{t_i \cdot x_i}}}{\sigma^2}\right)$$

4.4.4.1 *Window-based Classifier with Clip-based Output*

The output of the single-frame classifier is then filtered by the window-based classifier, to then generate video clips of all the windows that are contiguous or overlapping.

6 A GT video-clip is a sequence of contiguous frames where the pedestrian is present in the camera field of view.

7 Notice that although it's the same number as GT bounding boxes, this is a coincidence, and that 155 of these 1182 detections are FP.

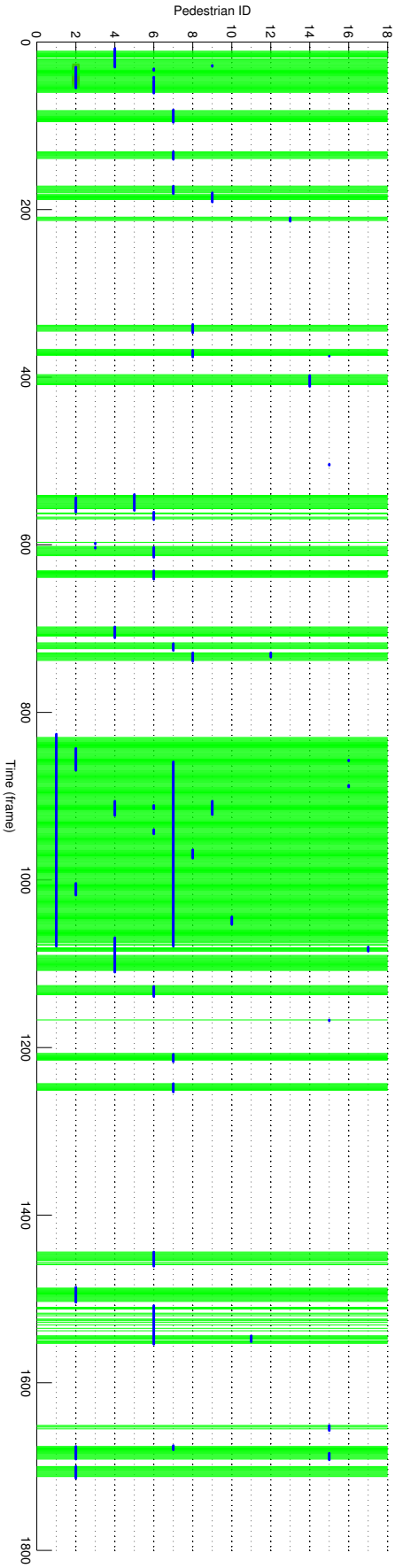


Figure 37: In blue dots we have the distribution of all 17 pedestrian appearances throughout the probe video sequence. In green vertical lines we have all the detections provided by the PD.

Scenario	GT	Occ Filt	FP class	Dets	MDs	FPs
$\text{MANUAL}_{\text{all}}$	1	NA	NA	1097	0	0
$\text{MANUAL}_{\text{clean}}$	1	ON GT	NA	416	681	0
$\text{MANUAL}_{\text{cleanhalf}}$	1	ON GT	NA	208	889	0
DIRECT	0	OFF	OFF	964	288	155
FPCLASS	0	OFF	ON	964	288	155
FPOCC30	0	ON 30%	ON	854	362	119

Table 14: In this table we summarize the details of each scenario. **GT** indicates the use of ground truth for detection in a scenario, namely the hand-labeled BBs. When **GT** is set to 0 this means an automatic pedestrian detector is being used for detections. NA indicates that the use of the Occlusion Filter (**Occ Filt**) or the FP class is not applicable to experiments with GT data. The total number of detections and the amount of corresponding false positives which are passed to the RE-ID module are listed under **Dets** and **FPs**, respectively. The total number of missed detections are listed under **MDs**.

4.4.5 *Evaluation Metric*

The necessary GT for evaluating the RE-ID task is obtained by processing the original GT annotations, and the detections generated by the PD module. Each detection is associated with the label of a person or the special label for the FP class. The assignment is done associating each detection with the label of the GT BB that has the most overlap with it. The Pascal VOC criterion [41] is used to determine FPs: when the intersection between a detection BB and the corresponding BB from the original GT is smaller than half the union of the two, the detection is marked as a FP.

To perform the evaluation of all the window-based classifier with clip-based output experiments the metrics described in the previous section are computed: precision in frames ($\mathbf{Prec_f}$) and recall in all frames ($\mathbf{Rec_f}$). The Cumulative Matching Characteristic curve (CMC) is also computed for the single-frame classifier.

4.4.6 *Experiments*

Six scenarios were devised to illustrate the different aspects of the integrated PD and RE-ID system.

For each scenario, all parameters of our window-based classifier were varied, in the following range: $r \in [1,5]$, $d \in [1,20]$ and $w \in [1,1740]$, which adds up to 174 000 experimental runs for each scenario. Note that d=1, w=1 corresponds to the single-frame classifier.

In scenario MANUAL_{all} RE-ID is performed on all pedestrian appearances no matter how occluded or truncated they may be in the image. For all pedestrians there are some frames in which they are significantly truncated

	Best $\mathcal{T}$ (r,	d,	w)	Prec_f (%)	Rec_f (%)	Med r	Med d	Med w
Prec_f	(1,	$\geqslant 18$,	[18 20])	100	$\leqslant 5.7$	1	13	16
Rec_f	(5,	1,	$\geqslant 198$)	$\leqslant 0.9$	100	5	1	248

Table 15: Triplets of parameters that maximize each metric separately, on the $\text{MANUAL}_{\texttt{all}}$ scenario (described in the text), along with their respective metric values. The median value of each parameter for the 100 best triplets for each metric is also shown. From this it's observable that recall pulls for large rank r and threshold d=1, while precision prefers rank r=1, large d and pulls for the smallest possible window size w (w must always be $\geqslant$ to d).

(when they are entering or leaving the camera's field of view). These instances should be impossible for the RE-ID classifier to correctly classify, yet, this scenario provides a meaningful baseline for recall because there are absolutely no MDs. Note that this method of operation is not applicable in a real-world situation, since it requires manual annotation of every person in the video sequence.

In the $\text{MANUAL}_{\texttt{clean}}$ scenario RE-ID is performed on the 416 GT BBs where the pedestrians are fully visible, consistently with the *modus operandi* of the state of the art. This means that the RE-ID module works with unoccluded persons and BBs that are correctly centered and sized. Note that this method of operation is also not applicable in a real-world situation, since it also requires manual annotation of every person in the video sequence. This scenario is a baseline for precision and accuracy.

In scenario $\text{MANUAL}_{\text{cleanhalf}}$ RE-ID is performed in half the samples of $\text{MANUAL}_{\text{clean}}$ randomly selected. This scenario is devised to highlight the effect of having many MDs, since only 208 of the total 1097 pedestrian appearances are used.

Then, in scenario DIRECT the performance of the system is analyzed resulting from the naive integration of the PD and RE-ID modules. Note that the 155 FPs generated by the detector will be impossible for the RE-ID to correctly classify, since they do not have a respective class in the gallery.

Afterwards, in the FPCLASS scenario, ON the FP class is turned ON to evaluate our approach to address detection false positives.

Finally, in scenario FPOCC30 the Occlusion Filter is turned ON with the overlap threshold set to 30%. It has been determined in [115] that 30% is the best value for this parameter in this dataset.

Table 14 summarizes the details of these six scenarios.

4.4.7 *Results*

This section presents the results obtained with the proposed architecture, following the six scenarios described above in Section 4.4.6 and summarized in Table 14. The results will be thoroughly discussed in the following section.

First, the CMC is computed for all six scenarios with the single-frame classifier (Figure 38).

Then, it's analyzed which combination of parameters maximize each metric individually, on the $\text{MANUAL}_{\text{all}}$ scenario. The set of 100 best combination of parameters $\mathcal{T}=(r, d, w)$ is taken for each metric, and present the best and the median value of each parameter for that set (Table 15).

Finally all the 174 000 experimental runs are computed for each of the six scenarios. Table 16 summarizes the results for six considered scenarios.

In Figure 39, its plotted in the $\mathbf{Prec_f}$ vs $\mathbf{Rec_f}$ space, one point per combination of parameters (triplet $\mathcal{T}=(r, d, w)$). Each point is colored with its respective **F-score** (from blue to dark red). Its also marked with a circle the point corresponding to the $\mathcal{T}$ that maximizes the score defined in Equation 13, and with a square the $\mathcal{T}$ that maximizes Equation 13 while setting d and w to 1 (the respective single-frame classifier).

	F-score (%)		**$\mathbf{Prec_f}$ (%)**		**$\mathbf{Rec_f}$ (%)**	
	(1,5,10)	(1,1*,1*)	(1,5,10)	(1,1*,1*)	(1,5,10)	(1,1*,1*)
$\text{MANUAL}_{\text{all}}$	33.5	26.7	36.2	27.1	31.3	26.3
$\text{MANUAL}_{\text{clean}}$	33.8	28.1	67.0	47.4	22.6	20.0
$\text{MANUAL}_{\text{cleanhalf}}$	19.2	15.5	77.4	44.6	10.9	09.4
DIRECT	34.6	25.6	39.5	27.5	30.8	23.9
FPCLASS	36.3	25.6	44.1	29.1	30.8	22.9
FPOCC30	38.9	27.0	53.8	34.3	30.4	22.2

Table 16: Results for three combination of parameters that provide the best results overall, and under (1,1*,1*) results for the corresponding best single-frame classifier (setting d=1 and w=1). The first conclusion is taken comparing the (1,1*,1*) column with the others, where its visible that the improvement provided by the window-based classifier is significant (reaching up to 11% in the **F-score**). The second conclusion comes from comparing the F-score values between different scenarios. The FPCLASS scenario consistently outperforms the DIRECT scenario and FPOCC30 consistently outperforms the other two. This supports the claims that adding a false positive class to the classifier helps deal with the false positives of the pedestrian detector, and that adding the occlusion filter to reject detections of occluded pedestrians ultimately also helps the overall re-identification system. The final conclusion comes from comparing $\text{MANUAL}_{\text{clean}}$ with $\text{MANUAL}_{\text{cleanhalf}}$, where the drastic drop in **F-score** due to the added MDs is clearly visible, while the CMC does not account for this effect.

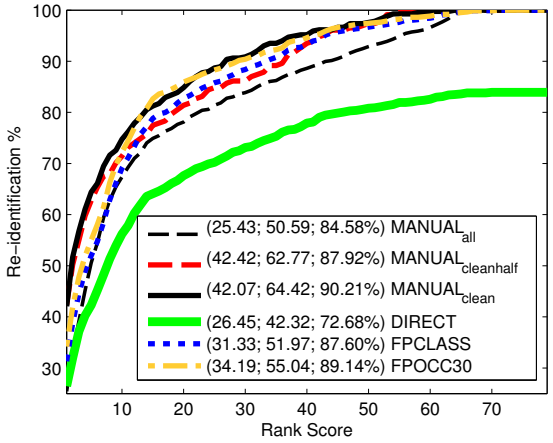


Figure 38: Cumulative Matching Characteristic curves comparing the performance of various configurations of the integrated Re-Identification system with a single-frame classifier. For details on the scenario of each experiment see Table 14. The three numbers for each line correspond to the first rank, fifth rank and normalized area under the curve respectively. Note how the MANUAL$_{cleanhalf}$'s CMC performance is roughly the same as the MANUAL$_{clean}$'s, highlighting how MDs are not penalized by the CMC metric.

4.4.8 *Discussion*

In this section it is discussed the results obtained by running the architecture in the scenarios described above.

Table 15 shows which combinations of parameters $\mathcal{T}$=(r,d,w) maximize each metric separately in the MANUAL$_{all}$ scenario. The triplets in the interval $\mathcal{T} = (1, \geqslant 18, [18\ 20])$ [8] maximize precision in frames. Rank r=1 and a large number of required detections d, and the smallest possible w, causes the highest possible confidence in each "detection" and produces video-clips with the least possible amount of false positives, thus optimizing precision. In this case, the parameters are so strict that only two video-clips are produced as output, and all frames of both video-clips contains their respective pedestrian, thus reaching 100% precision in frames. On the opposite side, all triplets in the interval $\mathcal{T} = (5, 1, \geqslant 198)$ maximize recall in frames. Large rank, which indicates small confidence in each detection, combined with a small number of required detections to present output, and a large window size w, causes the window-based classifier to capture almost everything. Therefore it does not miss any pedestrian appearance, and has 100% recall.

This experiment gives guidelines for the tuning of the parameters if one wishes to give more importance to precision or recall, given the application. If increased precision is desired, r should be reduced, d increased, while keeping w small. If someone wishes to maximize recall, he should increase r, reduce d and increase w. Note that in this last case, the amount of data shown to the operator is much larger, but a high-security application may require it.

8 Note that for a given $\mathcal{T}$, w needs to be always greater or equal to d

Now lets analyze Table 16 and compare the results between scenarios (rows) and between experiments in each scenario (columns). The first and foremost conclusion can be observed comparing the results for window-based classification with single-frame classification (first column of each metric against the column under $\mathcal{T}$=(1,1*,1*)). Window-based classification consistently outperforms single-frame classification, in all experiments, under the F-score metric defined in (13). This supports the claim that window-based classification improves results overall. The second important conclusion comes from comparing F-score values of the different scenarios. FPOCC30 consistently outperforms FPCLASS which consistently outperforms the experiments under the DIRECT scenario. This gives evidence that the proposed modules (FP class and occlusion filter) help deal with some of the issues of integrating the PD with RE-ID.

Let us now analyze each scenario individually. Scenario $\text{MANUAL}_{\text{all}}$ is one baseline, it has absolutely no MDs, thus it exhibits the best recall (see the first line of Table 16). The precision is low, because many instances of pedestrian appearances are truncated or occluded up to a point to make it difficult or even impossible for the single-frame classifier to correctly classify with rank r=1. This lowers the **F-score** and CMC performances.

Scenarios $\text{MANUAL}_{\text{clean}}$ and $\text{MANUAL}_{\text{cleanhalf}}$ other, complementary baselines. They sport the most number of MDs of all scenarios, thus exhibiting the lowest recall values. On the other hand, because they pass only the "clean" detections to the classifier, they achieve the lowest amount of mis-classifications and thus the highest precision values. Note how the CMC plot reports very good performances for both $\text{MANUAL}_{\text{clean}}$ and $\text{MANUAL}_{\text{cleanhalf}}$ (see Figure 38), while the **F-score** and recall values (see Table 16) clearly differentiate between the two scenarios: $\text{MANUAL}_{\text{cleanhalf}}$ achieves much worse **F-score** and recall than $\text{MANUAL}_{\text{clean}}$, due to the much higher number of MDs in the first scenario. These results show that the CMC plot is largely unaffected by different numbers of MDs and that precision and recall statistics provide complementary information to characterize the performance of integrated RE-ID systems.

In the DIRECT scenario, the naive integration of the PD and RE-ID exhibits the expected low performance (the lowest in Figure 38 and in F-score on Table 16). However, the best triplets of parameters $\mathcal{T}$=(r,d,w) are always low rank[9], in the region where the negative effect of not having a FP class is not particularly noticeable/relevant (see the first points of Figure 38). This makes results not that much worse than the rest of the scenarios. In the literature, FPs are either not considered to the classification, or their influence in the final performance is ignored. If indeed the FPs are considered, the CMC does not reach 100% (see green curve in Figure 38).

In the FPCLASS scenario the RE-ID module is able to classify a fraction of the FPs as such, therefore it exhibits better precision than DIRECT. The pedestrians that are wrongly classified as FPs won't decrease precision directly since

9 From the experiments conducted, it was observed that the best $\mathcal{T}$ always had rank r lower than 3.

the system does not measure precision of the FP class. However, they will decrease recall, because those instances are not recovered and shown to the user. Nevertheless, the loss of recall by this fact, is largely compensated by the improvement in precision in the window-based classifier results, and experiments in the FPCLASS scenario consistently outperform ones conducted in the DIRECT scenario, under the F-score metric. Note that, when comparing the DIRECT experiment with this one, It is visible that the CMC over-penalizes FPs. The area under the CMC is drastically smaller in the DIRECT experiment, while the **F-score** is just mildly inferior. This supports the assertion that it is of interest to complement the CMC with other metrics when integrating RE-ID with PD.

Finally, in scenario FPOCC30 it is confirmed that this operation mode is the best one. It consistently shows a better F-score performance, as well as precision. It is confirmed that applying the occlusion filter is a good compromise between having some MDs from the rejected detections and having a good re-identification performance, since it outperforms experiments from all other scenarios.

In Figure 39, where it's plotted all the 174 000 experimental runs for each scenario, one point per experiment, it is demonstrated the effectiveness of using a window-based classifier. All points in the figure indicates the performance of window-based classifiers with different combination of parameters, and the square indicates the performance of the respective single-frame classifier (with the best r) in that scenario. In all the six sub-figures (scenarios), the square (single-frame classification) is always surpassed by many possible window-base classifier parameter combinations. Also notice that the FPOCC30 scenario exhibits the best compromise of precision and recall overall.

Of interest is also noting that for all experiments, the best 100 triplets in **F-score** had all rank lower than 3. This suggests that only the lowest ranks matter for practical applications of the window-based and single-frame classifiers.

4.4.8.1 *Concluding Remarks*

In summary, the most important observations are that:

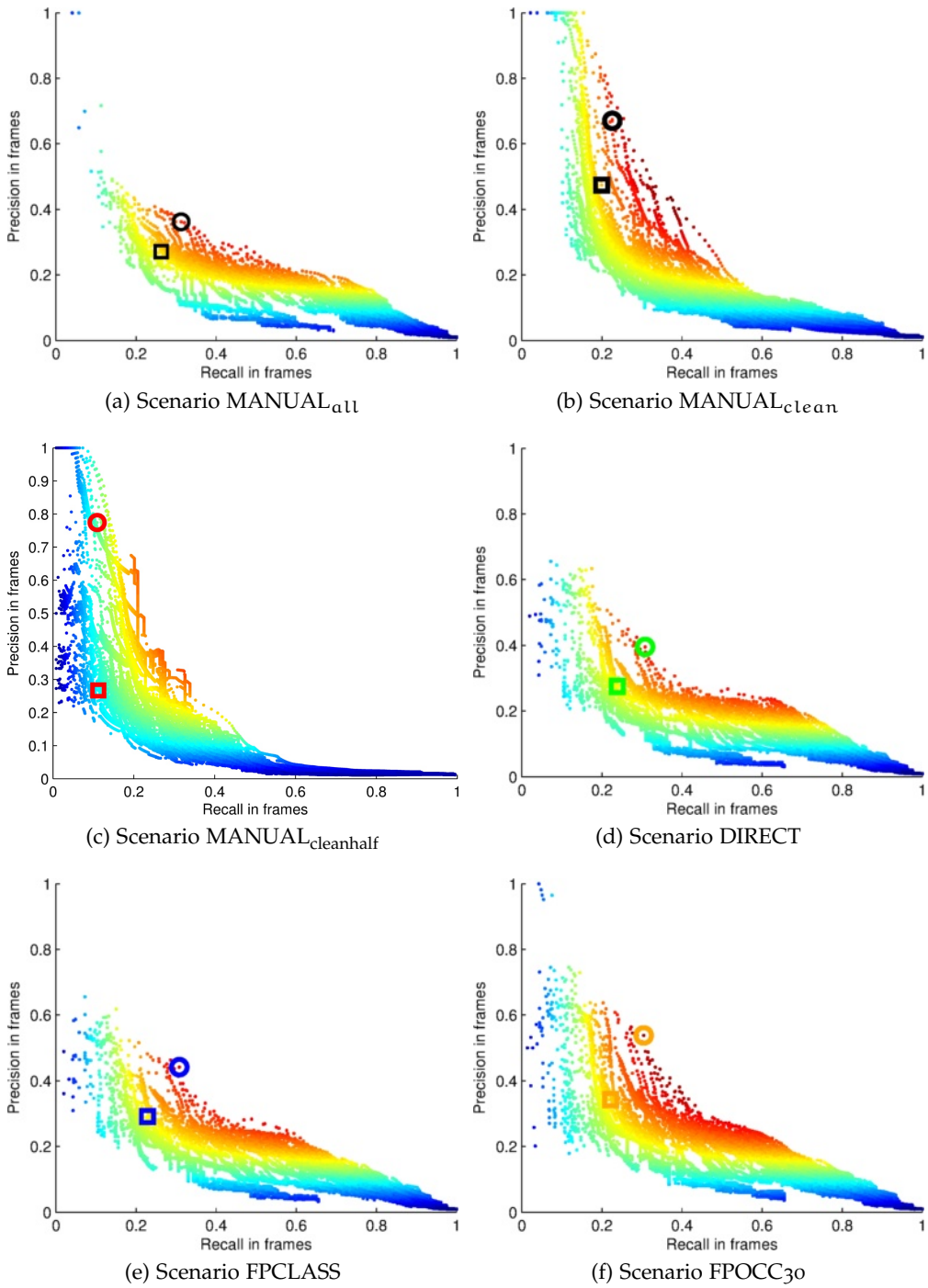
1. Window-based classification consistently outperforms single-frame classification, in all experiments, under the **F-score** metric. This means that window-based classification improves results overall.

2. The FPCLASS scenario outperforms the DIRECT scenario in both single-frame and window-based classification, for both CMC and **F-score** metric. This means that the FP class is an important module that should be used when integrating PD algorithms into the RE-ID pipeline.

3. The FPOCC30 scenario outperforms the FPCLASS scenario in both single-frame and window-based classification, for both CMC and **F-score** metric. This means that the occlusion filter is an important module that should be used when integrating PD algorithms into the RE-ID pipeline.

4. The sharp drop in **F-score** for the $\text{MANUAL}_{\text{cleanhalf}}$ scenario over the $\text{MANUAL}_{\text{clean}}$ scenario, while the CMC values remain mostly unchanged illustrate how the CMC does not penalize MDs.

5. The sharp drop in the CMC for the DIRECT scenario over the FPCLASS scenario, while the **F-score** is only a bit lower illustrate how the CMC overpenalizes FPs.



(a) Scenario MANUAL$_{all}$

(b) Scenario MANUAL$_{clean}$

(c) Scenario MANUAL$_{cleanhalf}$

(d) Scenario DIRECT

(e) Scenario FPCLASS

(f) Scenario FPOCC30

Figure 39: Precision in frames versus recall in frames for all 174 000 combination of parameters r, d and w, in all five scenarios detailed in Table 14. Each point represents an experiment with a given combination of the parameters. Precision is displayed in the y axis, Recall in the x axis, and the respective F-score defined in (13) colors each point (from blue to red). The circle corresponds to the triplet $\mathcal{T} = (1, 5, 10)$ that is one that depicts the best performance all around. The square represents the point that maximizes (13) for d and w set to 1 (a single-frame classifier). By comparing the circle to the square, it is immediately evident that there is a large boost in performance from using a window-based classifier.

5 CONCLUSIONS

Re-identification has many challenging issues that result from the high variability of the people's appearance in the camera images due to different illumination, different clothes, occlusions, postures and camera's opto-electric characteristics and perspective effects. Furthermore, a real re-identification system requires automatic detection of the pedestrians which leads to several other issues that hinder re-identification: false positives, missed detections, unreliable bounding boxes, and detections of occluded pedestrians.

In this work the problem of re-identification was analyzed in a holistic fashion. Despite a lot still has to be done for dealing with the high-variability of person's appearances, I was able to enhance the state-of-the-art both in feature extraction and classification. For feature extraction, following recent paradigms of part-based object representations, human detection was divided in body parts, to be able to extract features from semantically meaningful local regions. For classification a semi-supervised multi-view classification algorithm was used, which copes well with a small number of training samples and leverage unlabeled test data, whose sample size is often larger than the labeled test data. Moving towards the automation of re-identifications systems, I looked into the problems arising from using pedestrian detection algorithm for selecting image regions for re-identification. This brings problems due to unreliable bounding boxes, false positive and false negative detections. By letting body-part detection take care of the unreliability of detection bounding boxes, since detecting body-part locations corrects misalignments in the person detection; by training a false positive class to capture false positives of the detector, these can be handled by the classifier which couldn't before; by using an occlusion filter to prune out some detections of occluded pedestrians, re-identification performance increases, by removing hard to re-identify samples; and by using a window-based classifier to exploit the temporal coherence of pedestrian appearances, it filters out some spurious false positives and re-captures some missed detections. One other point of contribution was the proper evaluation of re-identification systems by proposing metrics that assess the impact of false positives and missed detections in the overall system to complement the usual metric employed by the re-identification community (CMC curves). Finally, classified pedestrians are tracked across cameras by the state-of-the-art Multiple Hypothesis Tracker.

Both the advances in feature extraction and in the integration of automatic detection with re-identification are general enough to be able to be applied to virtually any work in the literature. I expect these contributions to be widely used and boost research in integrated pedestrian detector and re-identification systems, bringing them closer to reality.

5.1 FUTURE WORK

Finally we discuss possible future avenues of work towards real-life application in person re-identification problems.

This thesis illustrated how Multi-View consistently improves re-identification performance over the other tested methods. However, many other degrees of freedom in the multi-view formulation are still open. In this section we provide a few points worth exploring for eventual performance gains.

To make the solution to the problem closed form, we averaged the estimated labels for each view via the matrix C. Optimization can be attempted not only over the function space – that yields the functions that project the feature space into the label space – but also over the $\mathbf{C}$ matrix (that integrates the contributions from each feature).

Also, it is assumed from the results that each single feature classifier trained during Multi-View is better than the same single feature classifier when trained alone alone, but still inside the multi-view framework. Nevertheless it would be interesting to run experiments that make this explicit.

In [113] they allow for one parameter per view in the regularization terms that govern the *approximation error* ($\gamma_1 \|f^1\|^2_{H^1} + \gamma_2 \|f^2\|^2_{H^2}$). To reduce the number of free parameters here it was opted for only one γ_A for all views ($\gamma_A \|f\|^2_K$). It would be interesting to study the effect of having one parameter per view to govern the *approximation error*, to see if further performance gains can be attained.

Finally, another avenue of interesting research is the analysis of how the parameters (r, d, w) of the window-based classifier vary given different base accuracy of the used single-frame classifier. What will be a good base value for those parameters for any classifier, or how should they be tuned given an expected accuracy of the single-frame classifier.

5.2 PUBLISHED WORKS

In the beginning of this thesis the issue of detecting and separately classifying people and robots was explored [45, 96], under the context of a multi-robot and camera test network [16].

Then, a new dataset (HDA dataset) was developed, from where to detect pedestrians [116] and benchmark algorithms. Following [116], and directly relevant to this thesis, an extension to the HDA dataset was proposed [47], termed herein HDA+, that added evaluation software specially tailored for re-identification. Both the dataset[1] and software[2] are available online.

Afterwards, [115] proposes and analyzes the use of a PD algorithm to provide detections to the RE-ID stage and [48] further extends the analysis with evaluation metrics that take into account the problems introduced by the non-ideal nature of automatic pedestrian detectors (Sections 3.1, 3.3.2 and 3.3.3).

1 You may request to download the HDA dataset at vislab.isr.ist.utl.pt/hda-dataset/

2 You can download the evaluation software and extras directly at github.com/vislab-tecnico-lisboa/hda_code

Further on, a semantic division of a pedestrian detection from where to extract descriptive features was developed [44] (Sections 3.1.1 and 3.2). After several baselines for re-identification were implemented and tested, I focused on a semi-supervised formulation for classification, that successfully fuses any number of different features – Multi-View [46] (Section 3.3.1). Finally, the re-identification work was integrated into a overarching tracking system that allows the correction of mistaken classifications, *i.e.*, when people change clothes – the Multiple Hypothesis Tracker [6] (Section 3.4).

A

APPENDIX DATA LABELLER

We made good use of Dollár's annotation tool [36, 35, 37]. But it was insufficient for our needs, therefor we improved upon it (improvement described in Figure 40)[1].

Figure 40: Note the small button on the top left of the panel - that is our addition. After having a video loaded, and a labeling file open, one click of that button extracts all detections into a separate folder: crops every detection from every frame, names them with the respective label and frame number, and stores them into a preteremined folder.

The challenge of improving this data labeling tool is well summarized in the very comments by the original author (Dollár) at the beginning of the code:

```
% The code below is fairly complex and poorly documented.
% Please do not email me with question about how it works.
```

A more extended summary would describe the lack of indentation and the use of lines of code like this:

```
for i=1:5, pLf.btn(i)=uic(pLf.h,btnPr{:},o4,'CData',icn5{i}); end
```

1 The improved version of the annotation tool is available on request.

B

APPENDIX XING METRIC LEARNING

In this appendix we describe in more detail the metric learning algorithm by Xing. Xing *et al.* [123] wishes to learn a metric of the following form:

$$d_{ML}(x,y) = \|x-y\|_A = \sqrt{(x-y)^T A (x-y)},$$

B.1 DEFINITIONS:

- Data $= \{x_1, \ldots, x_N\} \in \mathbb{R}^{N \times d}$
- $d^{ij} \in \mathbb{R}^d : d^{ij} = x_i - x_j$
- S: set of similar pairs.
- D: set of dissimilar pairs.

B.2 DIAGONAL A

In the simpler case where the goal is only to learn a diagonal matrix, the optimization can be simplified into the simpler form below:

$$\begin{aligned}
\min_A \quad & \sum_{(i,j)\in S} \|x_i - x_j\|_A^2 = \\
& = \sum_{(i,j)\in S} {d^{ij}}^T A\, d^{ij} = \sum_{(i,j)\in S} \sum_{k=1}^{d} {d^{ij}}_k^2\, A_{kk} = \\
& = \sum_{(i,j)\in S} w^T \cdot [{d^{ij}}_1^2\ {d^{ij}}_2^2 \ldots {d^{ij}}_d^2]^T \\
& = w^T \sum_{(i,j)\in S} [{d^{ij}}_1^2\ {d^{ij}}_2^2 \ldots {d^{ij}}_d^2]^T \qquad (14) \\
\text{s.t.} \quad & \sum_{(i,j)\in D} \|x_i - x_j\|_A = \sum_{(i,j)\in D} \sqrt{w^T \cdot [{d^{ij}}_1^2\ {d^{ij}}_2^2 \ldots {d^{ij}}_d^2]^T} \geqslant t \qquad (15) \\
& A \succcurlyeq 0 \qquad (16)
\end{aligned}$$

where $w = [A_{11}\ A_{22} \ldots A_{dd}]^T$.

B.3 FULL A

When optimizing over the full matrix A, Xing posed the problem in the following way:

$$\max_{A} \quad \sum_{(i,j)\in D} \|x_i - x_j\|_A = \sum_{(i,j)\in D} \sqrt{d^{ij\,T} A\, d^{ij}} \tag{17}$$

$$\text{s.t.} \quad \sum_{(i,j)\in S} \|x_i - x_j\|_A^2 = \sum_{(i,j)\in S} d^{ij\,T} A\, d^{ij} \leqslant t \tag{18}$$

$$A \succcurlyeq 0 \tag{19}$$

Note that restriction (18) can be rewritten as $w^T a \leqslant t$, considering and taking advantage of:

$$t = \sum_{k=1}^{d} \sum_{(i,j)\in S} d_k^{ij} d_k^{ij} / 1000 = \sum_{(i,j)\in S} \|d^{ij}\|_2^2 / 1000$$

$$a = \left[A^{11}\, A^{12} \dots A^{1d}\, A^{21} \dots A^{dd}\right]^T$$

$$W_{kl} = \sum_{(i,j)\in S} d_k^{ij} \cdot d_l^{ij}$$

$$w = \sum_{(i,j)\in S} \left[d_1^{ij} d_1^{ij}\; d_1^{ij} d_2^{ij} \dots d_1^{ij} d_d^{ij}\; d_2^{ij} d_1^{ij} \dots d_d^{ij} d_d^{ij}\right]^T$$

$$w1 = w/\|w\|, \quad t1 = t/\|w\|$$

where t is a scalar, a is the unrolled matrix of A, w is also an unrolled matrix, but of $W \in \mathbb{R}^{d\times d}$, and $w1$ is the normalization of w.

B.4 IMPLEMENTING IN CVX

This optimization problem can now be easily implemented in MATlab with CVX's optimization toolbox[1] as follows:

```
A = eye(d,d)*0.1;
t = w' * unroll(A)/100;
cvx_begin
cvx_quiet(false);
    variable a(length(unroll(A)));
    maximize(sum(sqrt(data_diff_unrolled*a)))
    subject to
        w*a <= t*ones(ts_dim,1);
        A>= 0
cvx_end
```

1 http://cvxr.com/cvx/

B.5 SPEEDED-UP CODE BY XING

The speeded-up code of the initial problem (Section B.3) by Xing goes like this[2]:

1. Given an A (initial, or resulting from a previous full iteration), the code iterates between obeying each constraint, until it gets to an A that obeys both:
   a) Given restriction (18) rewritten form $w^T a \leqslant t$, take $w^T a$, and compare it with t. If greater than t then step once in the direction of $w1$: $a = a + (t1 - w1^T \cdot a) \cdot w1$, where $t1 = t/\|w\|$;
   b) Satisfying restriction (19) is done simply by setting the negative eigenvalues of the resulting A to 0;
2. If both constraints are satisfied, then step in the gradient ascent of (17), and yield an A;
3. Return to step 1 if minimum not reached.

One trick used to speed up convergence is to not only compute the gradient of the objective function, but also compute the gradient of constraint (18), and then, taking from the objective function's gradient, only the orthogonal part to the constraint's gradient, taking a step in a direction that also 'minimizes' the disruption of constraint (18).

2 the matlab code of the speeded up version by Xing is available on request.

BIBLIOGRAPHY

[1] Mark Aaron Abramson. *Pattern search algorithms for mixed variable general constrained optimization problems*. PhD thesis, École Polytechnique de Montréal, 2002.

[2] 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*, 28(12):2037 –2041, dec. 2006. ISSN 0162-8828. doi: 10.1109/TPAMI.2006.244.

[3] Le An, M. Kafai, Songfan Yang, and B. Bhanu. Reference-based person re-identification. In *Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference on*, pages 244–249, 2013. doi: 10.1109/AVSS.2013.6636647.

[4] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Pictorial Structures Revisited: People Detection and Articulated Pose Estimation. *Computer Vision and Pattern Recognition*, 2009. URL http://www.gris.informatik.tu-darmstadt.de/~{}sroth/pubs/cvpr09andriluka.pdf.

[5] David Miguel Antunes, David Martins de Matos, and José Gaspar. A Library for Implementing the Multiple Hypothesis Tracking Algorithm. arXiv:1106.2263v1 [cs.DS], 2011.

[6] D.M. Antunes, D. Figueira, D.M. Matos, A. Bernardino, and J. Gaspar. Multiple hypothesis tracking in camera networks. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 367–374, 2011. doi: 10.1109/ICCVW.2011.6130265.

[7] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, pages 337–404, 1950.

[8] Tamar Avraham, Ilya Gurvich, Michael Lindenbaum, and Shaul Markovitch. Learning Implicit Transfer for Person Re-identification. In *ECCV*, 2012.

[9] S. Bak, E. Corvee, F. Brémond, and M. Thonnat. Person re-identification using Haar-based and DCD-based signature. In *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, pages 1 –8, 29 2010-sept. 1 2010. doi: 10.1109/AVSS.2010.68.

[10] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Multiple-shot human re-identification by mean riemannian covariance grid. In *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on*, pages 179–184, 2011. doi: 10.1109/AVSS.2011.6027316.

[11] Slawomir Bak, Guillaume Charpiat, Etienne Corvee, Francois Bremond, and Monique Thonnat. Learning to Match Appearances by Correlations in a Covariance Metric Space. In *ECCV*, 2012.

[12] Slawomir Bak, Etienne Corvee, Francois Bremond, and Monique Thonnat. Boosted human re-identification using riemannian manifolds. *Image and Vision Computing*, 30(6-7):443 – 452, 2012.

[13] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara. 3DPes: 3D People Dataset for Surveillance and Forensics. In *Proceedings of the 1st International ACM Workshop on Multimedia access to 3D Human Objects*, pages 59–64, Scottsdale, Arizona, USA, November 2011.

[14] Y. Bar-Shalom, F. Daum, and J. Huang. The probabilistic data association filter. *Control Systems, IEEE*, 29(6):82–100, 2009. ISSN 1066-033X. doi: 10.1109/MCS.2009.934469.

[15] Y. Bar-Shalom, F. Daum, and J. Huang. The probabilistic data association filter. *Control Systems Magazine, IEEE*, 29(6):82–100, 2009.

[16] Marco Barbosa, Alexandre Bernardino, Dario Figueira, José Gaspar, Nelson Gonçalves, Pedro U Lima, Plinio Moreno, Abdolkarim Pahliani, José Santos-Victor, M Spaan, et al. ISRobotNet: A testbed for sensor and robot network systems. In *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pages 2827–2833. IEEE, 2009.

[17] M. Bauml and R.Stiefelhagen. Evaluation of local features for person re-identification in image sequences. In *IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, pages 291–296, 2011.

[18] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino. Multiple-shot person re-identification by chromatic and epitomic analyses. *Pattern Recognition Letters*, 33(7):898–903, 2012.

[19] Loris Bazzani, Marco Cristani, and Vittorio Murino. Sdalf: Modeling human appearance with symmetry-driven accumulation of local features. In Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy, editors, *Person Re-Identification*, Advances in Computer Vision and Pattern Recognition, pages 43–69. Springer London, 2014. ISBN 978-1-4471-6295-7. doi: 10.1007/978-1-4471-6296-4_3. URL http://dx.doi.org/10.1007/978-1-4471-6296-4_3.

[20] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool. Pedestrian detection at 100 frames per second. *CVPR*, 2012.

[21] T.E. Boult, R.J. Micheals, Xiang Gao, and M. Eckmann. Into the woods: visual surveillance of noncooperative and camouflaged targets in complex outdoor settings. *Proceedings of the IEEE*, 2001.

[22] Sławomir Bąk and François Brémond. Re-identification by covariance descriptors. In Shaogang Gong, Marco Cristani, Shuicheng Yan, and

Chen Change Loy, editors, *Person Re-Identification*, Advances in Computer Vision and Pattern Recognition, pages 71–91. Springer London, 2014. ISBN 978-1-4471-6295-7. doi: 10.1007/978-1-4471-6296-4_4. URL http://dx.doi.org/10.1007/978-1-4471-6296-4_4.

[23] Ricardo Cabral, Fernando De la Torre, J Costeira, and Alexandre Bernardino. Matrix completion for weakly-supervised multi-label image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[24] Ricardo S Cabral, Fernando De la Torre, João P Costeira, and Alexandre Bernardino. Matrix completion for multi-label image classification. In *NIPS*, 2011.

[25] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom pictorial structures for re-identification. In *Proceedings of the British Machine Vision Conference*, pages 68.1–68.11. BMVA Press, 2011. ISBN 1-901725-43-X. http://dx.doi.org/10.5244/C.25.68.

[26] DongSeon Cheng and Marco Cristani. Person re-identification by articulated appearance matching. In Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy, editors, *Person Re-Identification*, Advances in Computer Vision and Pattern Recognition, pages 139–160. Springer London, 2014. ISBN 978-1-4471-6295-7. doi: 10.1007/978-1-4471-6296-4_7. URL http://dx.doi.org/10.1007/978-1-4471-6296-4_7.

[27] Mario Christoudias, Raquel Urtasun, and Trevor Darrell. Bayesian localized multiple kernel learning. Technical Report UCB/EECS-2009-96, EECS Department, University of California, Berkeley, Jul 2009. URL http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-96.html.

[28] Thomas H. Cormen, Clifford Stein, Ronald L. Rivest, and Charles E. Leiserson. *Introduction to Algorithms*. McGraw-Hill Higher Education, 2001.

[29] Etienne Corvee, Slawomir Bak, and Francois Bremond. People detection and re-identification for multi surveillance cameras. *VISAPP*, 2012.

[30] I. J. Cox and S. L. Hingorani. An efficient implementation and evaluation of Reid's multiple hypothesis tracking algorithm for visual tracking. In *Pattern Recognition, 1994. Vol. 1 - Conference A: Computer Vision Image Processing., Proceedings of the 12th IAPR International Conference on*, pages 437–442, 1994. doi: {10.1109/ICPR.1994.576318}.

[31] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. *CVPR*, 2005. doi: 10.1109/CVPR.2005.177. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1467360.

[32] R. Danchick and G. E. Newnam. Reformulating Reid's MHT method with generalised Murty K-best ranked linear assignment algorithm. *Radar, Sonar and Navigation, IEE Proceedings*, 153(1):13–22, 2006.

[33] Casia database. http://www.sinobiometrics.com.

[34] M. Dikmen, E. Akbas, T. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. In *ACCV*, 2011.

[35] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *CVPR*, 2009.

[36] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. In *PAMI*, volume 99, 2011. doi: http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.155.

[37] Piotr Dollár. Piotr's Image and Video Matlab Toolbox (PMT). http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html.

[38] Piotr Dollár, Serge Belongie, and Pietro Perona. The Fastest Pedestrian Detector in the West. *BMVC*, 2010. doi: 10.5244/C.24.68. URL http://www.bmva.org/bmvc/2010/conference/paper68/index.html.

[39] Piotr Dollár, R Appel, and W Kienzle. Crosstalk Cascades for Frame-Rate Pedestrian Detection. *ECCV*, 2012. URL http://vision.ucsd.edu/~pdollar/files/papers/DollarECCV2012crosstalkCascades.pdf.

[40] A. Ess, B. Leibe, and L. Van Gool. Depth and appearance for mobile scene analysis. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, 2007. doi: 10.1109/ICCV.2007.4409092.

[41] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman. The Pascal visual object classes (VOC) challenge. *IJCV*, 2010.

[42] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2010)*, San Francisco, CA, USA, 2010. IEEE Computer Society.

[43] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 2010. ISSN 1939-3539. doi: 10.1109/TPAMI.2009.167. URL http://www.ncbi.nlm.nih.gov/pubmed/20634557.

[44] Dario Figueira and Alexandre Bernardino. Re-Identification of Visual Targets in Camera Networks a comparison of techniques. In *ICIAR*, 2011.

[45] Dario Figueira, Plinio Moreno, Alexandre Bernardino, José Gaspar, and José Santos-Victor. Optical flow based detection in mixed human robot environments. *Advances in Visual Computing*, pages 223–232, 2009.

[46] Dario Figueira, Loris Bazzani, Minh Ha Quang, Marco Cristani, Alexandre Bernardino, and Vittorio Murino. Semi-supervised multi-feature learning for person re-identification. In *AVSS*, 2013.

[47] Dario Figueira, Matteo Taiana, Athira Nambiar, Jacinto Nascimento, and Alexandre Bernardino. The hda+ data set for research on fully automated re-identification systems. *ECCV Workshop*, 2014.

[48] Dario Figueira, Matteo Taiana, Jacinto Nascimento, and Alexandre Bernardino. Toward automatic video based re-identification: Problems, methods and evaluation techniques. *submitted to Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2014.

[49] François Fleuret, HoreshBen Shitrit, and Pascal Fua. Re-identification for improved people tracking. In Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy, editors, *Person Re-Identification*, Advances in Computer Vision and Pattern Recognition, pages 309–330. Springer London, 2014. ISBN 978-1-4471-6295-7. doi: 10.1007/978-1-4471-6296-4_15. URL http://dx.doi.org/10.1007/978-1-4471-6296-4_15.

[50] Itzhak Fogel and Dov Sagi. Gabor filters as texture discriminator. *Biological cybernetics*, 61(2):103–113, 1989.

[51] Per-Erik Forssén. Maximally stable colour regions for recognition and matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, USA, June 2007. IEEE Computer Society, IEEE.

[52] N. Gheissari, T. B. Sebastian, P. H. Tu, J. Rittscher, and R. Hartley. Person reidentification using spatiotemporal appearance. In *CVPR*, volume 2, pages 1528–1535, 2006.

[53] Andrew Gilbert and Richard Bowden. Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity. *Computer Vision–ECCV 2006*, pages 125–136, 2006.

[54] Andrew Gilbert and Richard Bowden. Incremental, scalable tracking of objects inter camera. *Comput. Vis. Image Underst.*, 111(1):43–58, jul 2008. ISSN 1077-3142. doi: 10.1016/j.cviu.2007.06.005. URL http://dx.doi.org/10.1016/j.cviu.2007.06.005.

[55] R Girshick, Pedro Felzenszwalb, and D McAllester. Object detection with grammar models. *PAMI*, 2011. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.231.2429&rep=rep1&type=pdf.

[56] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recongnition, reacquisition and tracking. In *IEEE Intl. Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, 2007.

[57] Doug Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *In IEEE International*

Workshop on Performance Evaluation for Tracking and Surveillance, Rio de Janeiro, 2007.

[58] Douglas 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.

[59] Donald Greenberg. Color spaces for computer graphics. In *in Computer Graphics (SIGGRAPH '78 Proceedings*, 1978.

[60] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *Distributed Smart Cameras, 2008. ICDSC 2008. Second ACM/IEEE International Conference on*, pages 1 –6, sept. 2008. doi: 10.1109/ICDSC.2008.4635689.

[61] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *Proc. of the IEEE Conf. on Distributed Smart Cameras*, volume 2, pages 1–6, 2008.

[62] M. T. Harandi, C. Sanderson, A. Wiliem, and B. C. Lovell. Kernel analysis over Riemannian manifolds for visual recognition of actions, pedestrians and textures. In *WACV*, 2012.

[63] Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Proc. Scandinavian Conference on Image Analysis (SCIA)*, 2011. The original publication is available at www.springerlink.com.

[64] Anil K Jain. *Fundamentals of digital image processing*, pages 68, 71, 73. Prentice-Hall, Inc., 1989.

[65] O. Javed, Z. Rasheed, K. Shafique, and M. Shah. Tracking across multiple cameras with disjoint views. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 952–957 vol.2, 2003. doi: 10.1109/ICCV.2003.1238451.

[66] Omar Javed, Zeeshan Rasheed, Khurram Shafique, and Mubarak Shah. Tracking across multiple cameras with disjoint views. *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 952–957, 2003. doi: {10.1109/ICCV.2003.1238451}.

[67] K. Jungling and M. Arens. Local feature based person reidentification in infrared image sequences. In *IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, pages 448–455, 2010.

[68] K. Jungling, C. Bodensteiner, and M. Arens. Person re-identification in multi-camera networks. In *CVPRW*, 2011.

[69] Svebor Karaman, Giuseppe Lisanti, AndrewD. Bagdanov, and AlbertoDel Bimbo. From re-identification to identity inference: Labeling consistency by local similarity constraints. In Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy, editors, *Person Re-Identification*, Advances in Computer Vision and Pattern Recognition, pages 287–307. Springer London, 2014. ISBN 978-1-4471-6295-7. doi: 10.1007/978-1-4471-6296-4_14. URL http://dx.doi.org/10.1007/978-1-4471-6296-4_14.

[70] DoHyung Kim, Jaeyeon Lee, Ho-Sub Yoon, and Eui-Young Cha. A non-cooperative user authentication system in robot environments. *Consumer Electronics, IEEE Transactions on*, 53(2):804–811, May 2007. ISSN 0098-3063. doi: 10.1109/TCE.2007.381763.

[71] V. Kovalev and S. Volmer. Color co-occurrence descriptors for querying-by-example. In *Multimedia Modeling, 1998. MMM '98. Proceedings. 1998*, pages 32 –38, oct 1998. doi: 10.1109/MULMM.1998.722972.

[72] Ryan Layne, Timothy Hospedales, and Shaogang Gong. Towards person identification and re-identification with attributes. In *ECCV*, 2012.

[73] Ryan Layne, TimothyM. Hospedales, and Shaogang Gong. Towards person identification and re-identification with attributes. In Andrea Fusiello, Vittorio Murino, and Rita Cucchiara, editors, *Computer Vision ECCV 2012. Workshops and Demonstrations*, volume 7583 of *Lecture Notes in Computer Science*, pages 402–412. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-33862-5.

[74] Ryan Layne, TimothyM. Hospedales, and Shaogang Gong. Attributes-based re-identification. In Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy, editors, *Person Re-Identification*, Advances in Computer Vision and Pattern Recognition, pages 93–117. Springer London, 2014. ISBN 978-1-4471-6295-7. doi: 10.1007/978-1-4471-6296-4_5. URL http://dx.doi.org/10.1007/978-1-4471-6296-4_5.

[75] Thomas Leung and Jitendra Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *Int. J. Comput. Vision*, 43(1):29–44, June 2001. ISSN 0920-5691. doi: 10.1023/A:1011126920638. URL http://dx.doi.org/10.1023/A:1011126920638.

[76] V. Leung, J. Orwell, and S.A. Velastin. Performance evaluation of re-acquisition methods for public transport surveillance. In *Control, Automation, Robotics and Vision, 2008. ICARCV 2008. 10th International Conference on*, pages 705–712, Dec 2008. doi: 10.1109/ICARCV.2008.4795604.

[77] Annan Li, Luoqi Liu, and Shuicheng Yan. Person re-identification by attribute-assisted clothes appearance. In Shaogang Gong, Marco

Cristani, Shuicheng Yan, and Chen Change Loy, editors, *Person Re-Identification*, Advances in Computer Vision and Pattern Recognition, pages 119–138. Springer London, 2014. ISBN 978-1-4471-6295-7. doi: 10.1007/978-1-4471-6296-4_6. URL http://dx.doi.org/10.1007/978-1-4471-6296-4_6.

[78] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *ACCV*, 2012.

[79] Zhe Lin and LarryS. Davis. Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance. In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Paolo Remagnino, Fatih Porikli, Jörg Peters, James Klosowski, Laura Arns, YuKa Chun, Theresa-Marie Rhyne, and Laura Monroe, editors, *Advances in Visual Computing*, volume 5358 of *Lecture Notes in Computer Science*, pages 23–34. Springer Berlin Heidelberg, 2008. ISBN 978-3-540-89638-8. doi: 10.1007/978-3-540-89639-5_3. URL http://dx.doi.org/10.1007/978-3-540-89639-5_3.

[80] Haibin Ling and K. Okada. Diffusion distance for histogram comparison. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 246 – 253, june 2006. doi: 10.1109/CVPR.2006.99.

[81] Chunxiao Liu, Shaogang Gong, Chen Change Loy, and Xinggang Lin. Person Re-identication: What Features Are Important? In *ECCV*, 2012.

[82] Chunxiao Liu, Shaogang Gong, ChenChange Loy, and Xinggang Lin. Person re-identification: What features are important? In *Computer Vision ECCV 2012. Workshops and Demonstrations*. Springer Berlin Heidelberg, 2012.

[83] Chunxiao Liu, Chen Change Loy, Shaogang Gong, and Guijin Wang. Pop: Person re-identification post-rank optimisation. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 441–448. IEEE, 2013.

[84] Chunxiao Liu, Shaogang Gong, and Chen Change Loy. On-the-fly feature importance mining for person re-identification. *Pattern Recognition*, 47(4):1602–1615, 2014.

[85] Chunxiao Liu, Shaogang Gong, ChenChange Loy, and Xinggang Lin. Evaluating feature importance for re-identification. In Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy, editors, *Person Re-Identification*, Advances in Computer Vision and Pattern Recognition, pages 203–228. Springer London, 2014. ISBN 978-1-4471-6295-7. doi: 10.1007/978-1-4471-6296-4_10. URL http://dx.doi.org/10.1007/978-1-4471-6296-4_10.

[86] Xiao Liu, Mingli Song, Dacheng Tao, Xingchen Zhou, Chun Chen, and Jiajun Bu. Semi-supervised coupled dictionary learning for person re-identification. In *Computer Vision and Pattern Recognition (CVPR), 2014*

IEEE Conference on, pages 3550–3557, June 2014. doi: 10.1109/CVPR.2014.454.

[87] Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. Text classification using string kernels. *J. Mach. Learn. Res.*, 2:419–444, March 2002. ISSN 1532-4435. doi: 10.1162/153244302760200687. URL http://dx.doi.org/10.1162/153244302760200687.

[88] Chen Change Loy, Tao Xiang, and Shaogang Gong. Time-delayed correlation analysis for multi-camera activity understanding. *Int. J. Comput. Vision*, 90(1):106–129, oct 2010. ISSN 0920-5691. doi: 10.1007/s11263-010-0347-5. URL http://dx.doi.org/10.1007/s11263-010-0347-5.

[89] AndyJinhua Ma and Ping Li. Semi-supervised ranking for re-identification with few labeled image pairs. In Daniel Cremers, Ian Reid, Hideo Saito, and Ming-Hsuan Yang, editors, *Computer Vision – ACCV 2014*, volume 9006 of *Lecture Notes in Computer Science*, pages 598–613. Springer International Publishing, 2015. ISBN 978-3-319-16816-6. doi: 10.1007/978-3-319-16817-3_39. URL http://dx.doi.org/10.1007/978-3-319-16817-3_39.

[90] Bingpeng Ma, Yu Su, and Frederic Jurie. Bicov: a novel image representation for person re-identification and face verification. In *BMVC*, 2012.

[91] Bingpeng Ma, Yu Su, and Frédéric Jurie. Discriminative image descriptors for person re-identification. In Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy, editors, *Person Re-Identification*, Advances in Computer Vision and Pattern Recognition, pages 23–42. Springer London, 2014. ISBN 978-1-4471-6295-7. doi: 10.1007/978-1-4471-6296-4_2. URL http://dx.doi.org/10.1007/978-1-4471-6296-4_2.

[92] N. Martinel and C. Micheloni. Re-identify people in wide area camera network. In *CVPR Workshops*, 2012.

[93] Michael McCahill and Clive Norris. Cctv in london. *Report deliverable of UrbanEye project*, 2002.

[94] A. Mignon and F. Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2666–2672, June 2012. doi: 10.1109/CVPR.2012.6247987.

[95] A. Mogelmose, C. Bahnsen, and T. B. Moeslung. Tri-modal person re-identification with RGB, depth and thermal features. In *IEEE WPBVS*, 2013.

[96] Plinio Moreno, Dario Figueira, Alexandre Bernardino, and José Santos-Victor. People and mobile robot classification through spatio-temporal analysis of optical flow. *International Journal of Pattern Recognition and Artificial Intelligence*, 29(06):1550021, 2015.

[97] Manuel Mucientes and Wolfram Burgard. Multiple Hypothesis Tracking of Clusters of People. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, pages 692–697, 2006.

[98] C. Nakajima, M. Pontil, and B. Heisele T. Poggio. Full-body person recognition system. *Pattern Recognition*, 36(9):1997–2006, 2003.

[99] Songhwai Oh and Shankar Sastry. Tracking on a graph. In *IPSN '05: Proceedings of the 4th international symposium on Information processing in sensor networks*, page 26, 2005.

[100] Kenji Okuma, Ali Taleghani, Nando Freitas, James J. Little, and David G. Lowe. A boosted particle filter: Multitarget detection and tracking. In Tomas Pajdla and Jiri Matas, editors, *Computer Vision - ECCV 2004*, volume 3021 of *Lecture Notes in Computer Science*, pages 28–39. Springer Berlin Heidelberg, 2004. ISBN 978-3-540-21984-2. doi: 10.1007/978-3-540-24670-1_3. URL http://dx.doi.org/10.1007/978-3-540-24670-1_3.

[101] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton van den Hengel. Learning to rank in person re-identification with metric ensembles. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[102] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*, 2013.

[103] Leonid Pishchulin, Thorsten Thorm, and Max Planck. Articulated People Detection and Pose Estimation: Reshaping the Future. *CVPR*, 2012.

[104] Bryan Prosser, Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by support vector ranking. In *Proceedings of the British Machine Vision Conference*, pages 21.1–21.11. BMVA Press, 2010. ISBN 1-901725-40-5. doi:10.5244/C.24.21.

[105] Minh H. Quang, Loris Bazzani, and Vittorio Murino. A unifying framework for vector-valued manifold regularization and multi-view learning. In Sanjoy Dasgupta and David Mcallester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 100–108. JMLR Workshop and Conference Proceedings, May 2013. URL http://jmlr.csail.mit.edu/proceedings/papers/v28/haquang13.pdf.

[106] D. Ramanan, D. A. Forsyth, and A. Zisserman. Tracking people by learning their appearance. *Pattern Analysis and Machine Intelligence, IEEE*

Transactions on, 29(1):65 –81, jan. 2007. ISSN 0162-8828. doi: 10.1109/TPAMI.2007.250600.

[107] Donald B. Reid. An Algorithm for Tracking Multiple Targets. *IEEE Transactions on Automatic Control*, 24:843–854, 1979.

[108] Riccardo Satta, Federico Pala, Giorgio Fumera, and Fabio Roli. People search with textual queries about clothing appearance attributes. In Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy, editors, *Person Re-Identification*, Advances in Computer Vision and Pattern Recognition, pages 371–389. Springer London, 2014. ISBN 978-1-4471-6295-7. doi: 10.1007/978-1-4471-6296-4_18. URL http://dx.doi.org/10.1007/978-1-4471-6296-4_18.

[109] C. Schmid. Constructing models for content-based image retrieval. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 2, pages II–39 – II–45 vol.2, 2001. doi: 10.1109/CVPR.2001.990922.

[110] Cordelia Schmid. Constructing models for content-based image retrieval. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 2, pages II–39. IEEE, 2001.

[111] W.R. Schwartz and L.S. Davis. Learning discriminative appearance-based models using partial least squares. In *Computer Graphics and Image Processing (SIBGRAPI), 2009 XXII Brazilian Symposium on*, pages 322–329, 2009. doi: 10.1109/SIBGRAPI.2009.42.

[112] W.R. Schwartz and L.S. Davis. Learning Discriminative Appearance-Based Models Using Partial Least Squares. In *Proceedings of the XXII Brazilian Symposium on Computer Graphics and Image Processing*, 2009.

[113] Vikas Sindhwani and David S. Rosenberg. An rkhs for multi-view learning and manifold co-regularization. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 976–983, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390279. URL http://doi.acm.org/10.1145/1390156.1390279.

[114] Matteo Taiana, Jacinto Nascimento, and Alexandre Bernardino. An improved labelling for the INRIA person data set for pedestrian detection. *IbPRIA*, 2013.

[115] Matteo Taiana, Dario Figueira, Athira Nambiar, Jacinto Nascimento, and Alexandre Bernardino. Towards fully automated person re-identification. In *VISAAP*, 2014.

[116] Matteo Taiana, Athira Nambiar, Dario Figueira, Alexandre Bernardino, and Jacinto Nascimento. A multi-camera video data set for research on high-definition surveillance. *Int. Journal of Machine Intelligence and Sensory Signal Processing*, 2014.

[117] Luis F. Teixeira and Luis Corte-Real. Video object matching across multiple independent views using local descriptors and adaptive learning. *Pattern Recogn. Lett.*, 30(2):157–167, January 2009. ISSN 0167-8655. doi: 10.1016/j.patrec.2008.04.001. URL http://dx.doi.org/10.1016/j.patrec.2008.04.001.

[118] DungNghi Truong Cong, Catherine Achard, Louahdi Khoudour, and Lounis Douadi. Video sequences association for people re-identification across multiple non-overlapping cameras. In Pasquale Foggia, Carlo Sansone, and Mario Vento, editors, *Image Analysis and Processing – ICIAP 2009*, volume 5716 of *Lecture Notes in Computer Science*, pages 179–189. Springer Berlin Heidelberg, 2009. ISBN 978-3-642-04145-7. doi: 10.1007/978-3-642-04146-4_21.

[119] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 606–613, 2009. doi: 10.1109/ICCV.2009.5459183.

[120] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *Proc. European Conference on Computer Vision (ECCV)*, 2014.

[121] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *ICCV*, 2007.

[122] Xiaogang Wang and Rui Zhao. Person re-identification: System design and evaluation overview. In Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy, editors, *Person Re-Identification*, Advances in Computer Vision and Pattern Recognition, pages 351–370. Springer London, 2014. ISBN 978-1-4471-6295-7. doi: 10.1007/978-1-4471-6296-4_17. URL http://dx.doi.org/10.1007/978-1-4471-6296-4_17.

[123] Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart Russell. Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems 15*, pages 505–512. MIT Press, 2002.

[124] Jingjing Yang, Yuanning Li, Yonghong Tian, Lingyu Duan, and Wen Gao. Group-sensitive multiple kernel learning for object categorization. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 436–443, 2009. doi: 10.1109/ICCV.2009.5459172.

[125] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking: A survey, 2006.

[126] Y. Zhang and S. Li. Gabor-lbp based region covariance descriptor for person re-identification. In *In Proc. of Int. Image and Graphics Conf.*, pages 368–371, 2011.

[127] W. Zheng, Shaogang Gong, and T. Xiang. Associating groups of people. In *BMVC*, 2009.

[128] W. Zheng, S. Gong, and T. Xiang. Re-identification by relative distance comparison. *IEEE Trans. Pattern Anal. Mach. Intell.*, PP(99):1, 2012. ISSN 0162-8828. doi: 10.1109/TPAMI.2012.138.

[129] Wei-Shi Zheng. Transfer re-identification: From person to set-based verification. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '12, pages 2650–2657, Washington, DC, USA, 2012. IEEE Computer Society. ISBN 978-1-4673-1226-4. URL http://dl.acm.org/citation.cfm?id=2354409.2354973.

[130] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by probabilistic relative distance comparison. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 649–656, June 2011. doi: 10.1109/CVPR.2011.5995598.

[131] Wei-Shi Zheng, Shaogang Gong, , and Tao Xiang. Re-identication by Relative Distance Comparison. *PAMI*, 2012.

[132] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Group association: Assisting re-identification by visual context. In Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy, editors, *Person Re-Identification*, Advances in Computer Vision and Pattern Recognition, pages 183–201. Springer London, 2014. ISBN 978-1-4471-6295-7. doi: 10.1007/978-1-4471-6296-4_9. URL http://dx.doi.org/10.1007/978-1-4471-6296-4_9.