A Window-Based Classifier for Automatic Video-Based Reidentification

Dario Figueira, Matteo Taiana, Jacinto C. Nascimento, *Member, IEEE*, and Alexandre Bernardino, *Member, IEEE*

***Abstract*—The vast quantity of visual data generated by the rapid expansion of large scale distributed multicamera networks, makes automated person detection and reidentification (RE-ID) essential components of modern surveillance systems. However, the integration of automated person detection and RE-ID algorithms is not without problems, and the errors arising in this integration must be measured (e.g., detection failures that may hamper the RE-ID performance). In this paper, we present a window-based classifier based on a recently proposed architecture for the integration of pedestrian detectors and RE-ID algorithms, that takes the output of any bounding-box RE-ID classifier and exploits the temporal continuity of persons in video streams. We evaluate our contributions on a recently proposed dataset featuring 13 high-definition cameras and over 80 people, acquired during 30 min at rush hour in an office space scenario. We expect our contributions to drive research in integrated pedestrian detection and RE-ID systems, bringing them closer to practical applications.**

***Index Terms*—Camera networks, pedestrian detection (PD), reidentification (RE-ID), video surveillance.**

I. Introduction

THIS paper addresses the problem of person reidentification (RE-ID) in camera networks. Given a set of pictures of previously observed persons, a practical RE-ID system must locate and recognize such people in the stream of images flowing from the camera network. In this paper, we consider a set of cameras with low overlap covering our research institute facilities, as a representative example of real-world office-spaces.

In a space with access control, when a person enters or passes some key locations where identity verification is possible (e.g., with face recognition or access control), pictures are acquired and stored in a gallery, associated to the corresponding person's identity. Such pictures are then used to recognize the person as it passes in different points of the camera network, or query about its location in segments of the recorded videos. A RE-ID system can thus be used to search and track a particular person on the network, which are capabilities of great interest for research on several topics of interest, such as modeling activities, mining physical social networks and human-robot interaction.

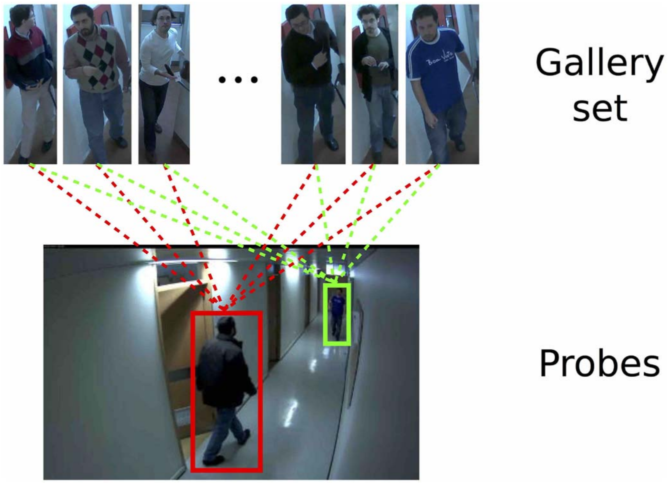


Fig. 1. Typical RE-ID algorithm is based on a gallery set: a database that contains the persons to be reidentified at evaluation time. People detected in other images (probes) are matched to such database with the intent of recognizing their identities. Classically, RE-ID algorithms are evaluated with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

Most current research on RE-ID focuses on the recognition problem using evaluation datasets where images of persons are manually prepared and labeled. The appearance information of persons stored in the gallery is matched against manually cropped images of persons in the network cameras (probe data); see Fig. 1 for an illustration of the process. However, in most RE-ID applications of interest, it is necessary to detect the probe person's bounding box (BB) in an automated way. This is commonly accomplished using background subtraction and pattern recognition-based approaches. We focus on methods based on the latter, because of their applicability to a wider range of scenarios, including moving/shaking cameras, single frame images, and scenarios with other classes of moving agents (e.g., animals, cars, and robots).

Both in background subtraction and in pattern recognition methods, the output of pedestrian detection (PD) is subject to several forms of "noise," in particular missed detections (MDs), false positives (FPs) and BBs misaligned

Manuscript received October 28, 2015; revised January 16, 2016, March 16, 2016, and May 1, 2016; accepted June 30, 2016. Date of publication September 27, 2016; date of current version November 15, 2016. This work was supported in part by the Portuguese Government under Grants PEst-OE/EEI/LA0009/2013, PTDC/EIA-EIA/105062/2008, and CMUP-ERI/HCI/0046/2013. This paper was recommended by Associate Editor K. Huang.

The authors are with the Institute for Systems and Robotics, Department of Electrical and Computer Engineering, Instituto Superior Técnico, 1049-001 Lisboa, Portugal (e-mail: dfigueira@isr.ist.utl.pt).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMC.2016.2609153

with the detected people. When a RE-ID algorithm is combined with an automatic PD algorithm, the performance of the former will suffer from the imperfections of the output of the latter. Furthermore, it is worth noticing that a class of correct detections by the PD is particularly difficult for a RE-ID algorithm to handle: the detections of partially occluded people. Currently, very few works address these problems.

To leverage the combination of PD and RE-ID algorithms in a fully automated RE-ID system, in a previous work [30] we proposed two important extensions (see Section III). On one hand, the PD output is used to filter detections with a large degree of occlusion (Section III-A), which most likely contain a large amount of data not related to the detected person. Doing so prevents misidentifications at the RE-ID stage. On the other hand, the RE-ID module is trained to directly represent a class of FPs commonly detected in the camera network (Section III-B). This minimizes the false associations of probe noise to actual persons in the gallery.

In this paper, we propose improvements on the ability of an RE-ID algorithm to tackle errors on PD by filtering the RE-IDs in temporal windows, extending our previous architecture [30]. Additionally, we concatenate the contiguous positive time windows in video-clips rather than individual images, which permits a faster browsing and inspection of the output. The proposed method can be used with any other BB RE-ID algorithm.

II. Related Work

In the last years, research in RE-ID has spawn in two main directions: 1) the definition of features that allow discrimination between different persons in different cameras and 2) matching or classification techniques that make the use of features to answer to user queries in the video data.

Concerning the features used, a recent study has been conducted to evaluate the performance of local appearance features for pedestrian RE-ID [2]. The conclusion is that, conventional detectors, such as Harris and Stephens [15], and Mikolajczyk and Schmid [23], perform equally well, and that the SIFT descriptor outperformed the alternatives. Other approaches propose the combination of color and texture features (see [34]). Alternative methods comprising symmetric-driven accumulation of local features [35] or BiCov descriptor [21] that combines Gabor filter with covariance descriptors are also available in the literature. Contextual knowledge from surrounding people using human signature [35], or attribute weighting features [20] have been demonstrated to be useful. Features can be computed in static parts of the detection window [4], [12] or in dynamic regions located through body parts detectors, as proposed in [3].

Once feature descriptors have been extracted from the gallery and probe data, RE-ID can be formulated as a *matching* or *data association* problem. Some works use off-the-shelf classifiers applied directly to the extracted features, e.g., support vector machines or boosting [13], [27], [36]. Another direction is concerned with learning task-specific distance functions with metric algorithms [16], [18], [19], [22], [36]. For all these methods, training samples with identity labels is mandatory. In this line of research, large-margin nearest neighbor with rejection is proposed [6] to learn the most suitable metric for matching data from two distinct cameras, while other metrics include probabilistic relative distance comparison [36]. Rank-loss optimization has been used to improve accuracy in RE-ID [33] and a variant of locally preserving projections [14] is formulated over a Riemmanian manifold. Recently, pairwise constrained component analysis is proposed for metric learning tailored to address scenarios with a small set of examples [22]. Also [26] suggested a metric learning for RE-ID, where a local fisher discriminant analysis is defined by a training set.

Most of the aforementioned works use the VIPeR [13], ETHZ [8] and i-LIDS [35] data sets, focusing only on the matching problem, thus neglecting the automatic detection of people. Methods that actively integrate PD and RE-ID, or works that propose metrics to evaluate integrated RE-ID systems are still scarce in the literature. The work that relates the most with ours is that introduced in [24]. In that work, the system full flow (i.e., PD and RE-ID) is presented with a transient gallery to tackle open scenarios. They use RGB-D data which may be limiting in some environments. An additional approach that integrates PD and RE-ID is [17], in which infrared images are used from CASIA Gait database [5]. 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-ID performance is penalized when PD or tracking failures exist are not analyzed. The goal of this paper is precisely to explore how to enhance the link between PD and RE-ID algorithms to improve the overall performance. Although there exists some work concerning the proposal of metrics for performance evaluation of surveillance systems (i.e., [25] for far-field settings), this does not take into account the integration of PD and RE-ID.

III. Architecture for Automatic RE-ID

In this section we propose an architecture which integrates the PD and RE-ID stages. We propose methods to address different types of errors introduced by the automatic PD: 1) an occlusion filter to deal with occluded detections; 2) an FP class to deal with nonpeople detections; 3) a window-based classifier to exploit the temporal continuity of the pedestrians in the videos; and 4) providing output as video-clips to reduce the attentional load put on the user. The former two points were already presented in [30] and we revise them here for completeness. The latter two, are novel contributions of this paper. This architecture is fully illustrated in Fig. 2.

A. Occlusion Filter

The RE-ID performance can suffer by incorporating detections of occluded pedestrians. A BB including a person appearing under partial occlusion generates features different from a BB including the same person under full visibility conditions. When the partial occlusion is caused by a second person standing between the camera and the original person,

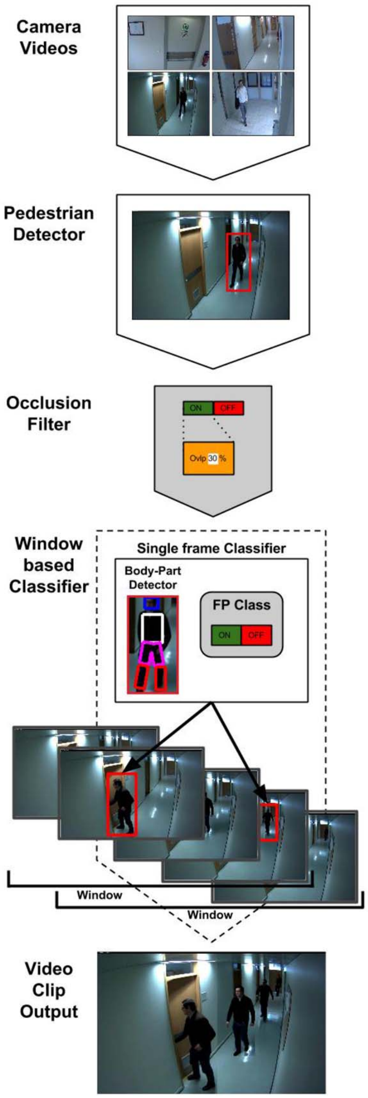


Fig. 2. Architecture of the proposed fully automated RE-ID system. The frames of the videos acquired by a camera network are processed by a PD algorithm to extract candidate BBs. The BBs are optionally processed by the occlusion filter. The RE-ID module computes the features corresponding to each BB and classifies it. The classification can optionally take into account a "false positive" class. The window-based classifier then takes the RE-IDs, and if there are enough positive RE-IDs in one or more temporal windows, outputs a video-clip with the combination of such windows.

the extracted features can be a mixture of those generated by the two people, making the identity classification especially hard [see illustration in Fig. 3 (top)]. For this reason, it would be advantageous for the RE-ID module to receive only BBs depicting fully visible people. Therefore, we devise the occlusion filter—a filtering block between the PD and the RE-ID modules (see Fig. 2)—with the intent of improving the RE-ID performance. The Occlusion Filter uses geometrical reasoning to reject BBs depicting partially occluded people.

Though the visibility information of which BB is occluding which is not available to the system, it can be estimated quite accurately with a scene geometry reasoning: in a typical scenario, where the camera is overhead, the camera's perspective projection makes proximal pedestrians extend to relatively lower regions of the image.

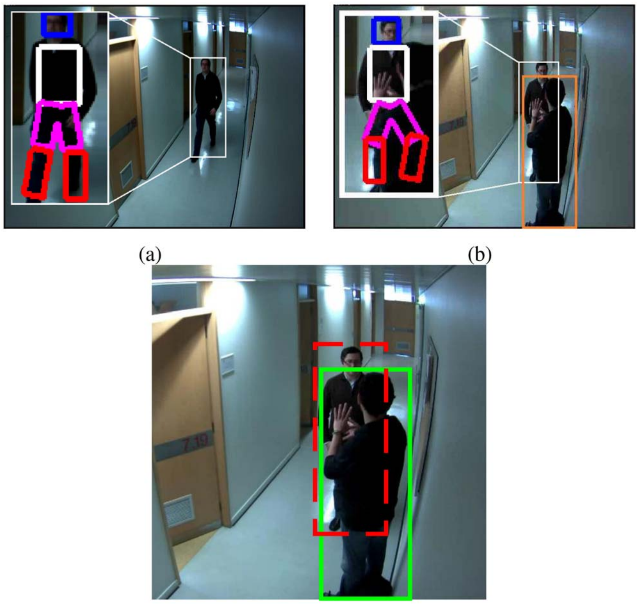

(a)
(b)

Fig. 3. (Top) Example of body-part detection for feature extraction in two instances. (a) Person appearing with full visibility and (b) under partial occlusion, with detected BBs overlap. The feature extraction on the occluded person mistakenly extracts some features from the occluding pedestrian. (Bottom) Example of geometrical reasoning. Two detection BBs overlap. The comparison between the lower sides of the two BBs leads to the conclusion that the person marked with the red, dashed BB is occluded by the person in the green, continuous BB.

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 Fig. 3 (bottom)]. 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. We define an overlap threshold for the filter, 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 our previous work [30], where we concluded that BBs should be rejected if more than 30% of their area is occluded.

B. False Positives Class

Pedestrian detector algorithms usually produce some FP detections. One way to improve their performance in a given context is to retrain the detector adding the FPs to the training as hard negatives on that specific context. However, some FPs may still remain, or the user may wish to use a *pretrained-off-the-shelf* PD algorithm where retraining is not an option. Therefore, another contribution of this paper is to adapt the RE-ID module to deal with the FPs produced by the PD. The standard RE-ID module cannot deal properly with FPs: each false positive (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 Fig. 4), we provide the RE-ID classifier with FP samples for it to be able to train a FP class. In these conditions a correct output for when a FP is presented on the RE-ID's

Fig. 4. Example FP samples in the FP class training set.

input exists: the FP class. Of course, introducing the FP class will also bring negative aspects; sometimes a valid person may be classified as a FP. We will see in the experiments section an analysis of this aspect.

C. Window-Based Classifier

Here, we describe the window-based classifier that exploits the temporal continuity of the pedestrians in the video to increase performance. It uses any BB classifier that gives a ranked output, and provides a binary output that informs if a given person is present in a continuous sequence of frames (window), without requiring tracking or data association between BBs in different frames.

At each frame the BB classifier provides a ranked list of RE-IDs for each BB. This list is denoted as *"BB reidentification ranked list."* If in this list there is an RE-ID of person X we say that there was a *"BB reidentification of X."*

We propose to represent RE-IDs in a per frame basis, instead of per BB. This is what will bring us the advantage of not requiring tracking or data association between BB along time. For each frame we compile the list of all BB RE-ID ranked lists which we denote as *"frame reidentification ranked list."* If in this list there is at least one RE-ID of person X in any BB, we say that there was a *"frame reidentification of X."*

In Fig. 5 we illustrate the distinction between BB RE-IDs and frame RE-IDs.

We recall that we do not make data association or tracking between BBs along frames, which distinguishes our algorithm from multishot approaches. For the same reason, other temporal filtering methods like the Hidden Markov Model [29] are not directly applicable.

In the following, we provide a definition of the main parameters involved in the window-based classifier that will allow us to tune its operation.

1) *Rank (r):* Given an ordered list of all matches for a probe sample against all classes in the gallery, rank denotes the largest index in the ordered list in which the correct match may show up for that sample It can also be used as a sensitivity parameter to set the algorithm operating point (e.g., high rank will improve recall and decrease precision).
2) *Window Size (w):* This stands for the number of frames of the window under consideration.
3) *Detection Threshold (d):* This variable controls the required minimum number of frame RE-IDs of rank r within a window of size w to grant a positive window.

Therefore, a positive window has w frames, containing at least d frame RE-IDs of rank r of a given person.

Algorithm 1 Window Classification Algorithm

```
INPUT: Query person X, Window W of size w, r, d
OUTPUT: $W_X$ (binary test of person X present in window)
for all frames $f$ in Window W do
    for all bounding boxes $b$ in $f$ do
        $L_b^f$ = ranked list of $b$ ▷ (bounding box ranked list)
    end for
    $L^f = \bigcup_b L_b^f$ ▷ (frame ranked list)
    if person X in $L^f$ with at least rank $r$ then
        $D_X^f = 1$ ▷ frame re-identification
    else
        $D_X^f = 0$
    end if
end for
$\#D_X = \sum_f D_f^X$
if $\#D_X$ then
    output: $W_X$ = true
else
    output: $W_X$ = false
end if
```

In Algorithm 1, we present the synopsis of the proposed window-based classifier.

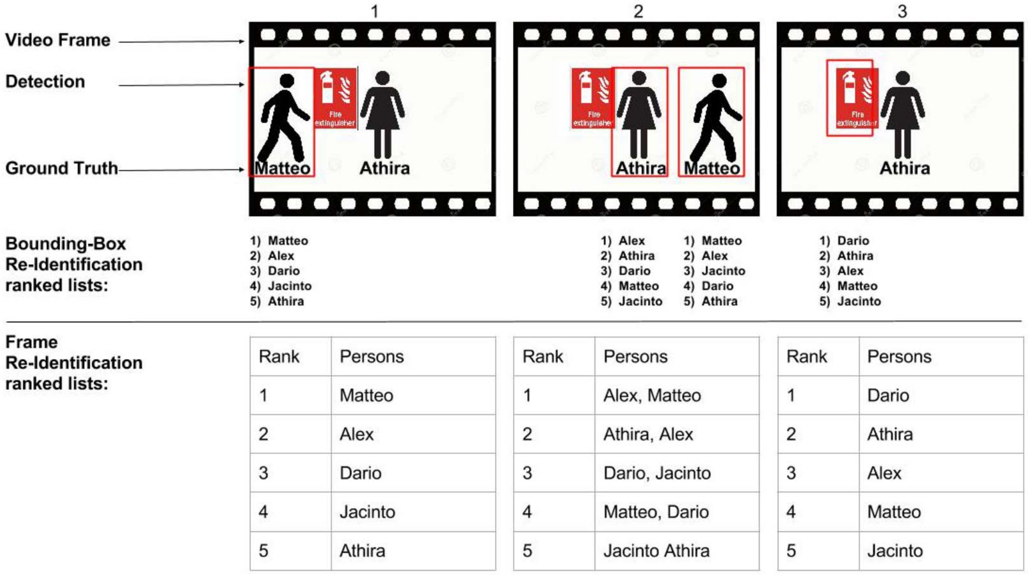
Based on the parameters just introduced above, we propose to use such triplet of parameters $\mathcal{T} = (r, d, w)$, to tune the algorithm's performance. For instance one detection with a corresponding RE-ID of rank 1 ($d = 1$ and $r = 1$) does not provide enough/reasonable confidence to justify giving output to an human operator. In fact, given the low average RE-ID rate of rank 1, it is required to have several rank 1 RE-IDs of that pedestrian, in a short period of time, to have a reasonable confidence that the pedestrian is indeed present. Therefore, we study the necessary rank r, size of window w and required number of detections d to optimize performance.

In Fig. 6, we illustrate the output of the window-based classifier for the different possible queries, for a few set of parameters, using as input data the data in Fig. 5. In the figure, we can see that increasing the temporal window size eventually removes spurious detections, and using a stricter rank r improves precision.

This filtering method will have at least two positive effects on RE-ID performance.

1) When the parameter d is >1 (minimum number of RE-IDs greater than 1 for a window to be provided as output) then spurious FPs and spurious misclassifications are always filtered out, thus improving precision.
2) When the parameter w is >2 (minimum width of the window greater than 2 then MDs or misclassifications that fall between d correct RE-IDs will always be recovered, since that whole window is provided as output, thus improving recall.

We stress that the proposed window-based classification works with any single-shot RE-ID algorithm, and does not



Bounding-Box Re-Identification ranked lists:

1	2	2	3
1) Matteo	1) Alex	1) Matteo	1) Dario
2) Alex	2) Athira	2) Alex	2) Athira
3) Dario	3) Dario	3) Jacinto	3) Alex
4) Jacinto	4) Matteo	4) Dario	4) Matteo
5) Athira	5) Jacinto	5) Athira	5) Jacinto

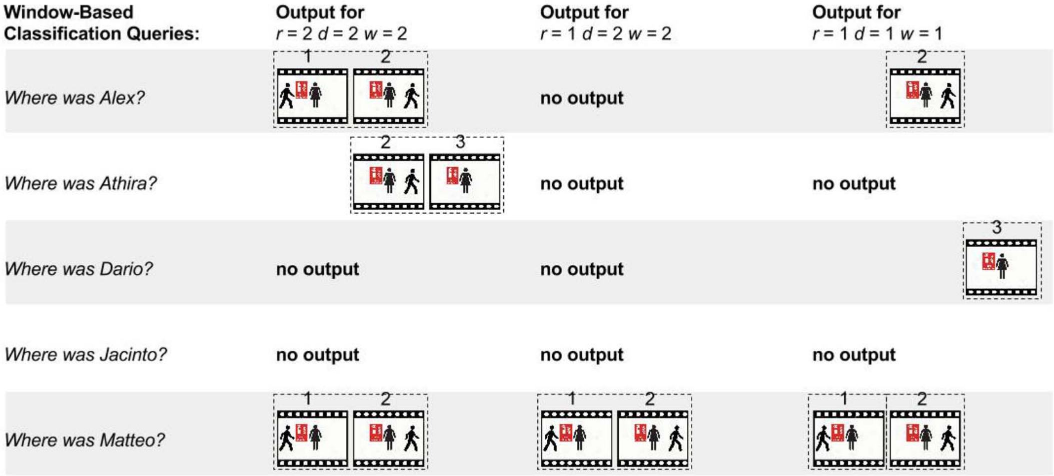
Frame Re-Identification ranked lists:

Rank	Persons
1	Matteo
2	Alex
3	Dario
4	Jacinto
5	Athira

Rank	Persons
1	Alex, Matteo
2	Athira, Alex
3	Dario, Jacinto
4	Matteo, Dario
5	Jacinto Athira

Rank	Persons
1	Dario
2	Athira
3	Alex
4	Matteo
5	Jacinto

Fig. 5. Example showing the BB RE-IDs versus frame RE-IDs.



Window-Based Classification Queries:	Output for $r = 2$ $d = 2$ $w = 2$	Output for $r = 1$ $d = 2$ $w = 2$	Output for $r = 1$ $d = 1$ $w = 1$
Where was Alex?	1 2	no output	2
Where was Athira?	2 3	no output	no output
Where was Dario?	no output	no output	3
Where was Jacinto?	no output	no output	no output
Where was Matteo?	1 2	1 2	1 2

Fig. 6. Example showing the window-based classifier output with different parameters, for the input data of Fig. 5.

require an in-camera tracker or manual selection of the images that belong to each individual, contrary to all of the works that do multishot RE-ID known to the authors.

D. Clip-Based Output

The output of our system is given in the form of video-clips that encapsulate frames with detections and RE-IDs of the person of interest, which decrease the attentional load of the user.

This is motivated by the following.

1) A human operator browses a couple of seconds of video (multiple frames) much faster than the equivalent number of individual frames.
2) A single detection and respective RE-ID does not guarantee a high degree confidence, therefore several of them are desirable to have higher confidence.
3) Pedestrian appearance in frames is not independent; they almost always appear in several contiguous frames.

Therefore, providing output in the form of video-clips, encapsulating several positive detections and RE-IDs of one given person, will facilitate the job of the human operator.

One video-clip is generated for the union of all positive windows of a certain pedestrian that overlap or are contiguous. In Fig. 2 we show an example with window size of $w = 4$, minimum number of detections of $d = 2$, and rank of $r = 1$. The person appears in four frames and is only detected and reidentified in two frames. Note how, albeit only being reidentified in two frames, the final output video-clip contains all four frames of interest.

IV. Metrics

The standard metric for 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-IDs at index i

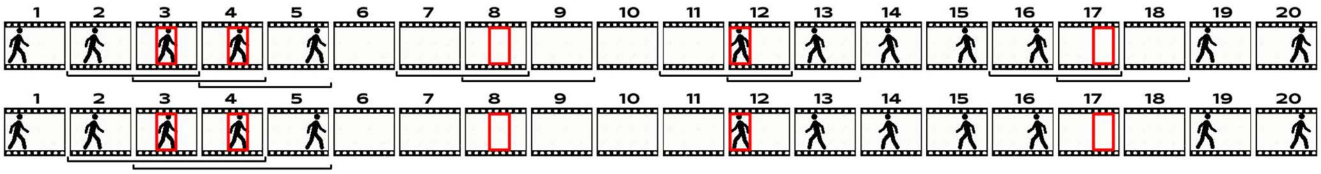


Fig. 7. Illustration of metrics calculation for a window-based classifier with $d = 1$ and $w = 2$ (top), and with $d = 2$ and $w = 3$ (bottom). In both, each read BB indicates RE-ID of rank $r = 1$. On the top, the pedestrian of interest appears in 12 frames of this video and the black brackets indicate the 13 frames that are shown as output of the system. From these 13 frames shown, seven of them truly contain the pedestrian of interest, therefore Prec_f is 7/13. From the 12 frames in which the pedestrian appears in the video, only seven are shown, thus Rec_f is 7/12. Note how although the detection and RE-ID in frame 17 is *erroneous* (a false-positive of the detector, and a lucky misclassification of the reidentifier), the corresponding video-clip shown *does indeed* contain the pedestrian of interest, and thus is a positive video-clip. In the bottom, the pedestrian also appears in 12 frames of this video and the black brackets indicate the 4 frames that are shown as output of the system. From these four frames shown, all four of them truly contain the pedestrian of interest, therefore Prec_f is 1. From the 12 frames in which the pedestrian appears in the video, only four are shown, thus Rec_f is 4/12.

in the ordered list of all matches for a probe sample against all classes in the gallery, then CMC is defined as

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

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

This means that when there are FP probes, without a FP class, each FP contributes to the denominator of (1) in the CMC calculation, but not to the numerator. Given the definition of $\text{ord}(i)$, and since the RE-ID classifier deals with each sample independently, then when there are FPs the number of correct RE-IDs is not affected. There simply are a greater number on incorrect RE-IDs, thus reducing every value of the CMC by the fraction of the amount of FPs relative to the total of probes. See (2) below for the mathematical representation of the original CMC equation when there are FPs

$$\text{CMC}'(r) = \sum_{i=1}^{r} \frac{\#\text{ord}(i)}{tp + \text{FP}} = \text{CMC}(r)\frac{tp}{tp + \text{FP}} < \text{CMC}(r). \tag{2}$$

When there are MDs, if on average, the samples missed are distributed proportionally to $\text{ord}(i)$, then the CMC does not change. See (4) for the mathematical representation of this. This means that the CMC does not penalize the MD introduced by the 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)

$$\text{CMC}''(r) = \sum_{i=1}^{r} \frac{\#\text{ord}(i) - \#\text{ord}(i)\frac{misseddetections(\text{MD}s)}{tp}}{tp - \text{MD}s} \tag{3}$$

$$= \sum_{i=1}^{r} \frac{\#\text{ord}(i)\left(1 - \frac{\text{MD}s}{tp}\right)}{tp - \text{MD}s} = \text{CMC}(r). \tag{4}$$

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-ID system.

In other fields such as object detection and tracking, precision and recall metrics are used to evaluate the algorithms.[1] 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, $i \in 1, \dots, P$, result in N^i presented videos v_n^i, $n \in 1, \dots, N^i$. 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.

1) *Precision in Frames (Prec_f):* Number of frames shown that do contain the pedestrian of interest over the total number of frames shown

$$\text{Prec}_f = \frac{1}{P}\sum_{i=1}^{P} \frac{\sum_{n=1}^{N^i} p\left(i, v_n^i\right)}{\sum_{n=1}^{N^i} t\left(v_n^i\right)}.$$

2) *Recall in Frames (Rec_f):* Number of frames shown that do contain the pedestrian of interest over the total number of frames in which the pedestrian appears

$$\text{Rec}_f = \frac{1}{P}\sum_{i=1}^{P} \frac{\sum_{n=1}^{N^i} p\left(i, v_n^i\right)}{gt(i)}.$$

See Fig. 7 for two illustrative examples and note the variation of the performance metrics in the same video for different d and w.

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) then precision should have more weight. Prec_f also penalizes positive video-clips that only have a few frames containing the pedestrian of interest, so it also limits the attentional load put on the user.

One of the possible combination is defined as follows:

$$\text{F} - \text{score} = 2 \cdot \frac{\text{Prec}_f \cdot \text{Rec}_f}{\text{Prec}_f + \text{Rec}_f}. \tag{5}$$

Here we compute the harmonic mean of Prec_f and Rec_f (usually called F-score). In this type of problems, there are usually

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

many more "true negatives" (frames without a given pedestrian) than false negatives, true positives and FPs combined. F-score, not only is a classical way to combine precision and recall, but also ignores the true negatives, being adequate to evaluate this problem.

Finally, since our window-based classifier has binary output, it cannot be evaluated by the CMC, it will only be evaluated by the precision and recall metrics. Nevertheless, since the CMC is the single most used metric in the RE-ID field, we felt compelled to include it in this paper and to stress how incomplete it is to evaluate a full PD+RE-ID system.

V. Experimental Setup

This section presents an extensive evaluation of the proposed system. The experimental evaluation will give emphasis to the novelties presented in the framework, namely: 1) the influence of the occlusion filter and the FP class presented in Section III and 2) the performance of our window-based RE-ID classifier for all the combination of parameters (r, d, w) comparing against the respective BB classifier $[\mathcal{T} = (r, 1, 1)]$, and baseline multishot methods assuming perfect tracking of pedestrians.

A. Materials

We work with a recently proposed high-definition data set [31].[2] The data was acquired using eight *high definition* cameras. Each sequence was acquired from a different camera and corresponds to 30 min of video during rush hour in our laboratory facilities. These eight video sequences contain 75 207 frames with 64 028 pedestrian appearances.

A set of images is collected beforehand and stored in a gallery associated to their identities. We use two disjoint sets for gallery and probes. More specifically, we select the best[3] images of seven out of the eight cameras sequences for the gallery, and use the left-out sequence as a probe set. The gallery is built by hand-picking one manually cropped BB 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 FP class (Section III-B) is built with the detections from the seven gallery sequences that have no overlap with any ground truth(GT) BB, for a total of 3972 detections in the FP class. In a realistic case, the system could be set to work on an automated by acquiring images early in 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 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 ground truth (GT) video-clips.[4] Fig. 8 displays in blue the appearances throughout the video of each of the 17 pedestrians.

B. Methods

1) Pedestrian Detection: In this paper, we used our implementation [32] of Dollár's [7] fastest pedestrian detector in the west (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[5] 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 assumptionl. This leads to the 964 elements that form the base set of detections, 155 of which are FPs. Fig. 8 displays in green the detections throughout the video of each of the 17 pedestrians.

2) Bounding-Box Reidentification: We use the state-of-the-art BB RE-ID algorithm from [11]. Given a BB provided by the GT or the PD module, the algorithm first detects body parts using pictorial structures [1]. The algorithm uses a body model consisting of head, torso, two thighs, and two shins during the detection phase, and then merges the areas corresponding to both thighs and both shins, respectively (see Fig. 3 for an example). Subsequently, the algorithm extracts color histograms and texture histograms from the distinct image regions. Finally it performs the classification based on the extracted features with the multiview algorithm [28].

3) Window-Based Classifier With Clip-Based Output: The output of the BB classifier is then filtered by the window-based classifier, to then generate video clips of all the positive windows that are contiguous or overlapping.

4) Multishot Methods: For comparison purposes, we have we implemented two multishot methods: the SDALF algorithm [10] and an implementation of a multishot wrapper of our multiview single shot classifier. In fact, this wrapper can be applied to any single shot classifier, as we explain below.

Multishot methods require tracklets (sequences of contiguous BBs of the same person) for RE-ID. We build these using the ground-truth data, as if a perfect tracker was available. Since our window-based classifier uses fixed parameters

[2] http://vislab.isr.ist.utl.pt/hda-dataset

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

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

[5] Notice that although this is the same number as GT BBs, this is a coincidence; 155 of these 1182 detections are FP.

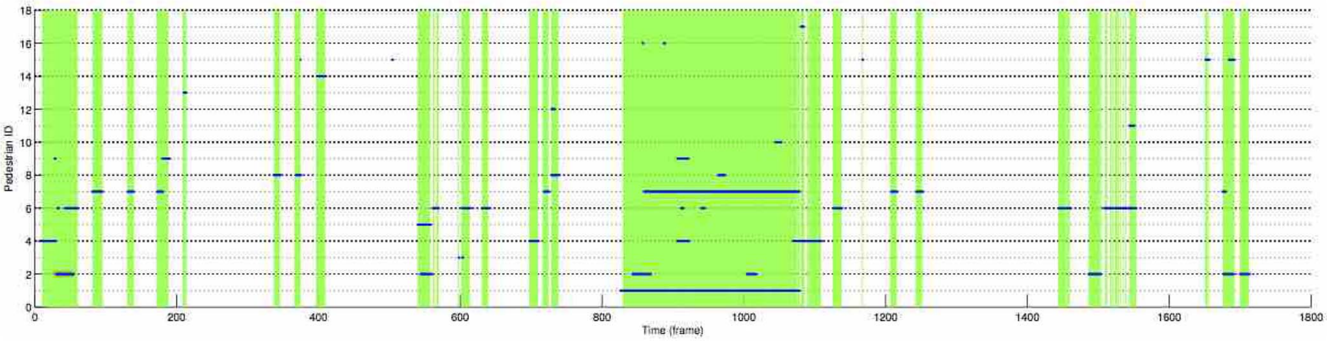


Fig. 8. In blue dots, we have the distribution of all 17 pedestrian appearances throughout the probe video sequence. In green vertical lines, we have all the detections provided by the PD.

for the window size (w) and the minimum number of detections (d), we also divide the tracklets into fixed size chunks of size m. Then, these chunks will be processed by a multishot method to provide the identities of each frame. Chunks with less than m BBs are discarded.

SDALF [10] is a state-of-the-art algorithm that computes stable color regions and recurrent motifs of high entropy in all BBs of the tracklet, weighted by body part. This descriptor is them matched to likewise descriptors, computed in the gallery, and outputs a ranked RE-ID list.

Our wrapper to multiview uses the confidence level of the rank 1 RE-IDs provided by the BB classifier, and outputs the identity of the person corresponding to the maximum confidence over all BBs of the tracklet.

C. Performance Evaluation

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 [9] 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 we compute the metrics described in the previous section, Prec_f and Rec_f. Since the CMC is the single most used metric in the RE-ID field, we fell compelled to include it in this paper and to stress how incomplete it is to evaluate a full PD+RE-ID system. We apply it to the BB classifier.

D. Scenarios

We devised six scenarios to illustrate the different aspects of the integrated PD and RE-ID system (see Table I).

For each scenario, we varied all parameters of our window-based classifier, 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 $d = 1$ and $w = 1$ corresponds to the BB classifier.

TABLE I
CONSIDERED EVALUATION SCENARIOS. SEE TEXT FOR EXPLANATION

Scenario	GT	Occ Filt	FP class	Dets	MDs	FPs
MANUAL$_{\text{all}}$	1	NA	NA	1097	0	0
MANUAL$_{\text{clean}}$	1	ON GT	NA	416	681	0
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

In scenario MANUAL$_{\text{all}}$ we perform RE-ID 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 (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 MANUAL$_{\text{clean}}$ scenario we perform RE-ID 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.

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

Then, in scenario DIRECT we analyze the performance of the system resulting from the naive integration of the PD and RE-ID modules. Note that the 155 FPs generated by the detector cannot be correctly reidentified, since they do not have a respective class in the gallery.

Afterward, in the FPCLASS scenario, we turn ON the FP class, thus providing the RE-ID algorithm with additional training samples (samples of FPs) to build an extra class (an FP class), and therefore evaluate our approach to address detection FPs.

TABLE II
TRIPLETS OF PARAMETERS THAT MAXIMIZE EACH METRIC SEPARATELY, ON THE MANUAL$_{\mathrm{ALL}}$ SCENARIO

	Best $\mathcal{T}$			**Prec$_f$**	**Rec$_f$**	**Median**		
	r,	d,	w	(%)	(%)	r	d	w
Prec$_f$	1,	≥ 18,	[18 20]	100	≤ 5.7	1	13	16
Rec$_f$	5,	1,	≥ 198	≤ 0.9	100	5	1	248

Finally, in scenario FPOCC30 we also turn ON the Occlusion Filter with the overlap threshold set to 30%. We have determined in our previous work [30] that 30% is the best value for this parameter in this dataset.

VI. Experimental Results

In this section, we discuss the results obtained by running our architecture in the scenarios described above.

Table II shows which combinations of parameters $\mathcal{T} = (r, d, w)$ maximize each metric separately in the MANUAL$_{\mathrm{all}}$ scenario. The triplets in the interval $\mathcal{T} = (1, \geq 18, [18\ 20])$[6] maximize Prec$_f$. 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 FPs, 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% Prec$_f$. On the opposite side, all triplets in the interval $\mathcal{T} = (5, 1, \geq 198)$ maximize Rec$_{\mathrm{f}}$. 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 allows us to have guidelines for the tuning of the parameters if we wish to give more importance to precision or recall, given the application. If we wish to increase precision we should reduce r, increase d, while keeping w small. If we wish to maximize recall, we 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.

We now analyze Table III 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 BB classification [first column of each metric against the column under $\mathcal{T} = (1, 1^*, 1^*)$]. Window-based classification consistently outperforms BB classification, in all experiments, under the F-score metric defined in (5). This supports our 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.

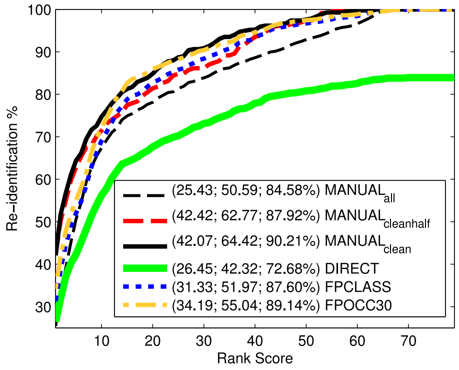


Fig. 9. CMCs comparing the performance of the BB classifier for various configurations of the integrated RE-ID system. The three numbers for each line correspond to the first rank, fifth rank, and normalized area under the curve, respectively.

Let us now analyze each scenario individually. Scenario MANUAL$_{\mathrm{all}}$ is one baseline, it has absolutely no MDs, thus it exhibits the best recall (see the first line of Table III). 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 BB classifier to correctly classify with rank $r = 1$. This lowers the F-score and CMC performances.

Scenarios MANUAL$_{\mathrm{clean}}$ and MANUAL$_{\mathrm{cleanhalf}}$ are also baselines, complementary to MANUAL$_{\mathrm{all}}$. They suffer from 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 misclassifications and thus the highest precision values. Note how the CMC plot reports very good performances for both MANUAL$_{\mathrm{clean}}$ and MANUAL$_{\mathrm{cleanhalf}}$ (see Fig. 9), while the F-score and recall values (see Table III) clearly differentiate between the two scenarios: MANUAL$_{\mathrm{cleanhalf}}$ achieves much worse F-score and recall than MANUAL$_{\mathrm{clean}}$, 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 Fig. 9 and in F-score on Table III). However, the best triplets of parameters $\mathcal{T} = (r, d, w)$ are always low rank,[7] in the region where the negative effect of not having a FP class is not particularly noticeable/relevant (see the first points of Fig. 9). 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 Fig. 9).

In the FPCLASS scenario the RE-ID module is able to classify a fraction of the FPs as such, therefore it exhibits better

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

[7] From the experiments conducted, we observed that the best $\mathcal{T}$ always had rank r lower than 3.

TABLE III
RESULTS FOR THE COMBINATION OF PARAMETERS THAT PROVIDE THE BEST RESULT OVERALL (1, 5, 10), AND UNDER (1, 1*, 1*) RESULTS FOR THE CORRESPONDING BB CLASSIFIER (SETTING $d = 1$ AND $w = 1$)

	F-score (%)		**Prec$_\mathrm{f}$ (%)**		**Rec$_\mathrm{f}$ (%)**	
	(1,5,10)	(1,1*,1*)	(1,5,10)	(1,1*,1*)	(1,5,10)	(1,1*,1*)
MANUAL$_\mathrm{all}$	33.5	26.7	36.2	27.1	31.3	26.3
MANUAL$_\mathrm{clean}$	33.8	28.1	67.0	47.4	22.6	20.0
MANUAL$_\mathrm{cleanhalf}$	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 IV
COMPARISON WITH MULTISHOT ALGORITHMS. MS5 AND MS10 ARE OUR MULTIVIEW METHOD WITH $w = 5$ AND $w = 10$ RESPECTIVELY. SDALF5 AND SDALF10 ARE THE SDALF ALGORITHM WITH 5 AND 10 SHOTS, RESPECTIVELY

	F-score (%)				**Prec$_\mathrm{f}$ (%)**				**Rec$_\mathrm{f}$ (%)**			
	MS10	SDALF10	MS5	SDALF5	MS10	SDALF10	MS5	SDALF5	MS10	SDALF10	MS5	SDALF5
MANUAL$_\mathrm{all}$	31.7	27.3	30.7	28.3	37.2	32.3	35.7	31.6	27.6	23.6	27.0	25.7
MANUAL$_\mathrm{clean}$	22.6	25.1	28.8	23.9	67.8	65.0	67.1	62.8	18.2	15.5	18.3	14.8
MANUAL$_\mathrm{cleanhalf}$	14.6	10.1	13.1	7.7	77.8	75.0	77.8	74.7	8.0	5.4	7.2	4.0
DIRECT	32.7	27.5	32.3	26.2	39.8	37.2	39.7	34.6	27.8	21.8	27.2	21.0
FPCLASS	34.8	29.4	32.6	27.7	44.2	41.3	44.0	40.3	28.7	22.8	25.9	21.1
FPOCC30	36.7	30.6	35.4	30.8	54.4	50.4	54.2	50.1	27.7	21.9	26.3	22.2

precision than DIRECT. The pedestrians that are wrongly classified as FPs will not decrease precision directly since 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, we see 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 our assertion that it is of interest to complement the CMC with other metrics when integrating RE-ID with PD.

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

In Fig. 10, where we plot all the 174 000 experimental runs for each scenario, one point per experiment, we demonstrate the effectiveness of using a window-based classifier. All points in the figure indicate the performance of window-based classifiers with different combination of parameters, and the square indicates the performance of the respective BB classifier in that scenario. In all the five sub-figures (scenarios), the square (BB 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.

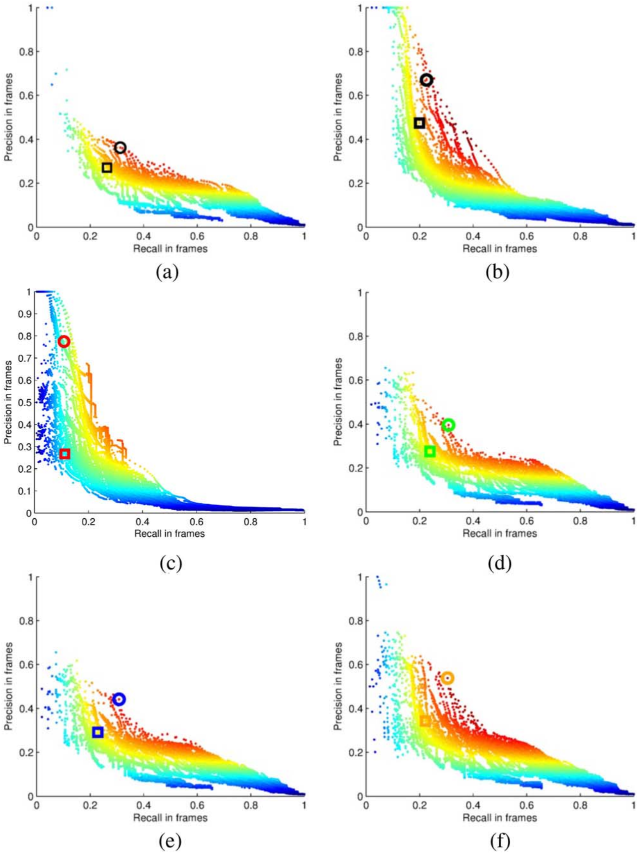
Of interest is also noting that for all experiments, the best 100 triplets in F-score had all rank lower than 3. This suggests that only the lowest ranks matter for practical applications of the window-based and BB classifiers.

In Table IV, we see the results of multishot algorithms in comparison with our window-based-classifier. The results suggest several assertions: the similar precision results for the window-based-classifier and our BB classifier wrapped by a 5-shot multishot wrapper indicate that both algorithms benefit from rejecting spurious wrong classifications, by requiring at least five classifications for a same person ID to provide output; the lower recall results of the multishot algorithms suggest that requiring tracklets of at least five sequential instances is strickter than our proposed window-based classifier. Our algorithm does not require tracks and only requires five loose frame RE-IDs to provide output, since the implemented multishot rejects more instances and has more MDs it has less recall. SDALF [10] has lower results overall which are probably due to its lower rank 1 BB RE-ID performance. Multishot's requirement of having tracklets of a given size may only be a plus when the detector+tracker can easily provide such sized tracklets, and/or when the BB classifiers do not have high-confidence wrong classifications. If the detector+tracker can easily provide the required sized tracklets, then the present multishot formulation will not cause too many MDs from smaller sized tracklets. If the BB classifiers have high-confidence wrong classifications, these classification doom whole tracklets to misclassification.

A. Concluding Remarks

In summary, the most important observations are as follows.

1) Window-based classification consistently outperforms BB classification, in all experiments, under the F-score



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

Fig. 10. 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 I. 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 (5) 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 (5) for d and w set to 1 (a BB 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. (a) Scenario MANUAL$_{\text{all}}$. (b) Scenario MANUAL$_{\text{clean}}$. (c) Scenario MANUAL$_{\text{cleanhalf}}$. (d) Scenario DIRECT. (e) Scenario FPCLASS. (f) Scenario FPOCC30.

metric. This means that window-based classification improves results overall.

2) The FPCLASS scenario outperforms the DIRECT scenario in both BB 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 BB 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 MANUAL$_{\text{cleanhalf}}$ scenario over the 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.
6) Window-based classification fares favorably against multishot due to being more relaxed in it is detection and tracking requirements which yields less MDs, and thus better recall.

VII. Conclusion

In this paper, we have taken an holistic view of the RE-ID problem and tackled several issues of interest for practical applications. First we define an architecture for a fully automatic system in closed-spaces, integrating a pedestrian detector with an RE-ID module. We propose, discuss and evaluate a few ways for “cleaning” the errors provided by the pedestrian detector, namely the FPs, the MDs and the occluded detections. A window-based classifier is proposed to exploit the continuity of persons in the video sequences. The window-based classifier takes noisy results from BB RE-ID classifiers and filters their errors through a voting mechanism. By merging overlapped windows into short video-clips, we are able to compact the information to be presented to the user. The parameters of the window-based classifier are analyzed by exhaustive experiments through precision/recall metrics that take into account the existence of MDs and FPs in the system, which are not completely assessed by CMC curves. These metrics provide a more complete view of the performance of the system, including the overhead required to a human operator on the verification of the systems output. Our window-based classificator fared well against multishot with a perfect tracker seemingly due to a combination of being more relaxed in its detection and tracking requirements and the BB classifiers poor rank 1 performance, which led to reduced MDs to our proposed algorithm.

References

[1] M. Andriluka, S. Roth, and B. Schiele, “Pictorial structures revisited: People detection and articulated pose estimation,” in *Proc. CVPR*, Miami, FL, USA, 2009, pp. 1014–1021.

[2] M. Bauml and R. Stiefelhagen, “Evaluation of local features for person re-identification in image sequences,” in *Proc. IEEE Int. Conf. Adv. Video Signal Based Surveillance*, Klagenfurt, Austria, 2011, pp. 291–296.

[3] L. Bazzani, M. Cristani, A. Perina, and V. Murino, “Multiple-shot person re-identification by chromatic and epitomic analyses,” *Pattern Recognit. Lett.*, vol. 33, no. 7, pp. 898–903, 2012.

[4] N. D. Bird, O. Masoud, N. P. Papanikolopoulos, and A. Isaacs, “Detection of loitering individuals in public transportation areas,” *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 2, pp. 167–177, Jun. 2005.

[5] *C Database*. Accessed on Jul. 1, 2015. [Online]. Available: http://www.sinobiometrics.com

[6] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, “Pedestrian recognition with a learned metric,” in *Proc. ACCV*, Queenstown, New Zealand, 2011, pp. 501–512.

[7] P. Dollár, S. J. Belongie, and P. Perona, “The fastest pedestrian detector in the west,” in *Proc. BMVC*, Aberystwyth, U.K., 2010, pp. 1–11.

[8] A. Ess, B. Leibe, and L. J. Van Gool, “Depth and appearance for mobile scene analysis,” in *Proc. ICCV*, Rio de Janeiro, Brazil, 2007, pp. 1–8.

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The Pascal visual object classes (VOC) challenge,” *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[10] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, “Person re-identification by symmetry-driven accumulation of local features,” in *Proc. CVPR*, San Francisco, CA, USA, 2010, pp. 2360–2367.

[11] D. Figueira *et al.*, “Semi-supervised multi-feature learning for person re-identification,” in *Proc. AVSS*, Kraków, Poland, 2013, pp. 111–116.

[12] N. Gheissari, T. B. Sebastian, and R. I. Hartley, "Person reidentification using spatiotemporal appearance," in *Proc. CVPR*, New York, NY, USA, 2006, pp. 1528–1535.

[13] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. ECCV*, Marseilles, France, 2008, pp. 262–275.

[14] 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 *Proc. WACV*, Breckenridge, CO, USA, 2012, pp. 433–439.

[15] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. Alvey Vis. Conf.*, Manchester, U.K., 1988, pp. 147–151.

[16] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *Proc. ECCV*, Florence, Italy, 2012, pp. 780–793.

[17] K. Jungling and M. Arens, "Local feature based person reidentification in infrared image sequences," in *Proc. IEEE Int. Conf. Adv. Video Signal Based Surveillance*, Boston, MA, USA, 2010, pp. 448–455.

[18] W. Li and X. Wang, "Locally aligned feature transforms across views," in *Proc. CVPR*, Portland, OR, USA, 2013, pp. 3594–3601.

[19] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Proc. ACCV*, Daejeon, South Korea, 2012, pp. 31–44.

[20] C. Liu, S. Gong, C. Loy, and X. Lin, "Person re-identification: What features are important?" in *Proc. ECCV Workshop*. Lecture Notes in Computer Science, vol. 7583. Florence, Italy, 2012, pp. 391–401.

[21] B. Ma, Y. Su, and F. Jurie, "BiCov: A novel image representation for person re-identification and face verification," in *Proc. BMVC*, Surrey, U.K., 2012, pp. 1–11.

[22] A. Mignon and F. Jurie, "PCCA: A new approach for distance learning from sparse pairwise constraints," in *Proc. CVPR*, Providence, RI, USA, 2012, pp. 2666–2672.

[23] K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector," in *Proc. ECCV*, Copenhagen, Denmark, 2002, pp. 128–142.

[24] A. Møgelmose, C. Bahnsen, T. B. Moeslung, A. Clapés, and S. Escalera, "Tri-modal person re-identification with RGB, depth and thermal features," in *Proc. IEEE WPBVS*, Portland, OR, USA, 2013, pp. 301–307.

[25] J. C. Nascimento and J. S. Marques, "Performance evaluation of object detection algorithms for video surveillance," *IEEE Trans. Multimedia*, vol. 8, no. 4, pp. 761–774, Aug. 2006.

[26] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local Fisher discriminant analysis for pedestrian re-identification," in *Proc. CVPR*, 2013, pp. 3318–3325.

[27] B. J. Prosser, W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," in *Proc. BMCV*, Aberystwyth, U.K., 2010, pp. 1–11.

[28] H. Q. Minh, L. Bazzani, and V. Murino, "A unifying framework for vector-valued manifold regularization and multi-view learning," in *Proc. ICML*, Atlanta, GA, USA, 2013, pp. 100–108.

[29] L. R. Rabiner and B.-H. Juang, "An introduction to hidden Markov models," *IEEE ASSP Mag.*, vol. 3, no. 1, pp. 4–16, Jan. 1986.

[30] M. Taiana, D. Figueira, A. Nambiar, J. Nascimento, and A. Bernardino, "Towards fully automated person re-identification," in *Proc. VISAAP*, 2014, pp. 140–147.

[31] M. Taiana, A. Nambiar, D. Figueira, A. Bernardino, and J. Nascimento, "A multi-camera video data set for research on high-definition surveillance," *Int. J. Mach. Intell. Sensory Signal Process.*, vol. 1, no. 3, 2014. doi: 10.1504/IJMISSP.2014.066428.

[32] M. Taiana, J. Nascimento, and A. Bernardino, "An improved labelling for the INRIA person data set for pedestrian detection," in *Proc. IbPRIA*, Funchal, Portugal, 2013, pp. 286–295.

[33] Y. Wu, M. Mukunoki, T. Funatomi, M. Minoh, and S. Lao, "Optimizing mean reciprocal rank for person re-identification," in *Proc. AVSS*, Klagenfurt, Austria, 2011, pp. 408–413.

[34] Y. Zhang and S. Li, "Gabor-LBP based region covariance descriptor for person re-identification," in *Proc. Int. Image Graph. Conf.*, Hefei, China, 2011, pp. 368–371.

[35] W.-S. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *Proc. BMVC*, London, U.K., 2009, pp. 1–11.

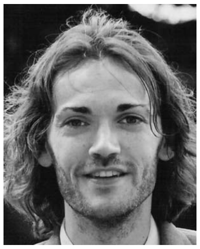
[36] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 653–668, Mar. 2013.

Dario Figueira received the master's degree in electrical and computer engineering from the Instituto Superior Técnico, Lisbon, Portugal, where he is currently pursuing the Ph.D. degree with the Computer and Robot Vision Laboratory, Institute for Systems and Robotics.

His current research interests include computer vision for video surveillance, reidentification, and pedestrian detection.

Matteo Taiana received the M.Sc. degree in computer engineering from Politecnico di Milano, Milan, Italy, in 2007, where he is currently pursuing the Ph.D. degree with the Computer and Robot Vision Laboratory, Institute for Systems and Robotics, Instituto Superior Técnico, Lisbon, Portugal.

His current research interest includes pedestrian tracking and more in general on computer vision for robotics.



Jacinto C. Nascimento (M'06) received the M.Sc. and Ph.D. degrees in electrical and computer engineering from the Instituto Superior Técnico, Lisbon, Portugal, in 1999 and 2006, respectively.

He is an Assistant Professor with the Department of Informatics and Computer Engineering, Instituto Superior Técnico, Lisbon, Portugal, where he is a Senior Researcher with the Computer and Robot Vision Laboratory, Institute for Systems and Robotics. His current research interests include image processing, pattern recognition, tracking, medical imaging, video surveillance, machine learning, and computer vision.

Alexandre Bernardino (M'06) received the M.Sc. and Ph.D. degrees in electrical and computer engineering from the Instituto Superior Técnico, Lisbon, Portugal, in 1997 and 2004, respectively.

He is an Associate Professor with the Department of Electrical and Computer Engineering, Instituto Superior Técnico, Lisbon, Portugal, where he is a Senior Researcher with the Computer and Robot Vision Laboratory, Institute for Systems and Robotics. His current research interests include the application of computer vision, machine learning, cognitive science, and control theory to advanced robotics and automation systems.