A Novel Human-Machine Hybrid Framework for Person Re-Identification from Full Frame Videos

Felix Oliver Sumari Huayta
Universidade Federal Fluminense,
Instituto de Computação,
Niteroi, Brazil, 24000-000,
Email: fsumari@id.uff.br

Esteban Gonzales Clúa
Universidade Federal Fluminense,
Instituto de Computação,
Niteroi, Brazil, 24000-000,
Email: esteban@ic.uff.br

Joris Guérin
Université de Toulouse,
Toulouse, France, 31000
Email: joris.guerin@laas.fr

***Abstract*—With the major adoption of automation for cities security, person re-identification (Re-ID) has been extensively studied. In this dissertation, we argue that the current way of studying person re-identification, i.e. by trying to re-identify a person within already detected and pre-cropped images of people, is not sufficient to implement practical security applications, where the inputs to the system are the full frames of the video streams. To support this claim, we introduce the Full Frame Person Re-ID setting (FF-PRID)[1] and define specific metrics to evaluate FF-PRID implementations. To improve robustness, we also formalize the hybrid human-machine collaboration framework, which is inherent to any Re-ID security applications. To demonstrate the importance of considering the FF-PRID setting, we build an experiment showing that combining a good people detection network with a good Re-ID model does not necessarily produce good results for the final application. This underlines a failure of the current formulation in assessing the quality of a Re-ID model and justifies the use of different metrics. We hope that this work will motivate the research community to consider the full problem in order to develop algorithms that are better suited to real-world scenarios.**

I. Introduction

In recent years, many security cameras were deployed in public places such as streets, malls or airports. Today, most of these video streams are monitored in real-time by security agents, which is expensive and rather inefficient as the amount of videos to analyze is tremendous. In contrast, automated video analysis [1] can process large amounts of videos simultaneously but is more prone to errors for complex tasks such as person re-identification [2]. In addition, even for automated video analysis systems, the final decision often rests with a human security agent, who triggers the appropriate actions. Hence, in practice it seems good to adopt hybrid approaches, where artificial intelligence models can screen the whole network in real time and select only relevant sequences for the monitoring agents.

Person Re-Identification (Re-ID) problem aims at searching a given person (query) in a network of non-overlapping cameras and raising an alert when this person appears in one of the video streams. It seeks to reproduce and enhance the human ability to recognize people in different scenarios, e.g. wearing different clothes, in a different pose, different illumination conditions, etc.

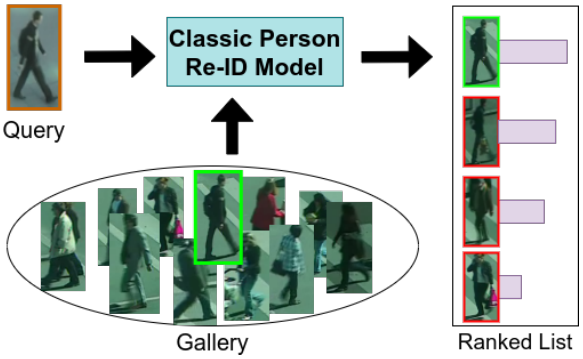


Fig. 1. Illustration of the Classic Person Re-ID (C-PRID) setting (Source: author).

The current formulation to address Re-ID is based on large databases of images representing human beings in a real-world environment [3]–[7]. These images are usually extracted using pedestrian detection models [8] and filtered manually to meet certain standards: each image should contain the entire body of exactly one person, centered and occupying most of the image(examples are shown on Fig. 1). From these datasets, a given image is selected as the query and the others constitute the search gallery. Then, the objective is to look for the query person within the gallery [2].

This approach is illustrated in Fig.1 where the output of a Classic Person Re-ID model is an ordered list with the most similar person on top. Sometimes, individual images are replaced by sequences of successive cropped images and the problem is called video-based Re-ID [9], [10]. From now on, the Re-ID setting considering pre-cropped images of persons as input is referred to as Classic Person Re-Identification (C-PRID). Recent successful methods to address C-PRID are mostly based on deep learning [11]–[15].

In practical, tasks of person re-ID system in video surveillance can be divided into three sub-modules [16]; (1)person detection ,(2)person tracking and (3)person retrieval. In general, the first two steps are investigated independently, so C-

[1]This work is based on an M.Sc. Dissertation of the first author Felix O. Sumari H.

Fig. 2. Full Frame Person Re-ID (FF-PRID) setting. (Source: author)

PRID works are focused on the last module in state of the art. Therefore, our motivation is to discuss the three modules as one task and solve the practical application problems.

Besides security applications [17]–[20], C-PRID is a useful building block for other practical applications such as 3D Multi Object-Tracking [21] or executing visual tasks for drones [22]. This work focuses on the practical security application, which consists of identifying in a network of non-overlapping cameras, a specific person being followed by human surveillance activity.

The main contributions of our work is the proposal of a new pipeline of the person Re-ID problem, called Full Frame Person Re-Identification (from now FF-PRID), which is better suited to implement and evaluate real-world security applications. By formalizing the natural collaboration occurring between an automated Re-ID system and the human monitoring agents, a hybrid and robust framework to address the FF-PRID problem is proposed, as well as two complementary metrics to assess the quality of any FF-PRID pipeline. Then, experiments are conducted to demonstrate the importance of considering the FF-PRID problem in its entirety.

II. Proposed Methodology

A. Full Frame Person Re-Identification(FF-PRID)

In short, in the FF-PRID setting, a successful model must analyze full frames to determine if the query is present in the stream, and if it is, when and where it appeared. The FF-PRID setting is illustrated in Fig. 2.

One can argue that the C-PRID problem can be easily derived from the FF-PRID setting by applying a pedestrian detection (PD) model [8] on the raw video stream, which is often done in practice. Indeed, some object detection models have demonstrated strong results for detecting human beings over the last few years [23], [24]. However, we argue that not considering the problem as a whole presents several issues:

1) The bounding boxes extracted by PD models may differ from the images in the reference datasets used for C-PRID training and evaluation, which have been filtered manually to only select clean images. This domain shift between the galleries used for training and the data encountered at inference time can decrease the quality of the model at run time, and thus induces a strong bias for model evaluation.
2) Even if both a good pedestrian detection model and a good Re-ID model are used, their small prediction errors might add up to produce poor overall results for the final application.
3) Not considering FF-PRID as an independent problem might dissuade the community from trying different approaches for the full application. Indeed, the vast availability of C-PRID datasets might take researchers away from trying other promising approaches such as end-to-end methods or video based methods, which have been shown to work for other computer vision problems [25], [26].
4) When developing a practical application, it is crucial to evaluate the quality of the entire pipeline before deploying it in production. To the best of our knowledge, frameworks and metrics to evaluate FF-PRID are missing in the literature.

B. A Human-Machine Hybrid Framework for FF-PRID

The classic formulation of person Re-ID consists in comparing a query image with all the images of a search gallery to output a set of similarity scores representing the Re-ID predictions. Conversely, this work considers the Full Frame Re-ID setting, which is better suited to implement and evaluate practical security applications. In this field, we introduce a hybrid framework, using human-machine collaboration to address the FF-PRID problem and we propose two new evaluation metrics to assess the quality of a FF-PRID model on a given dataset.

Framework. In the FF-PRID setting, the inputs to the system are a query image and a raw video from a security camera. Studying this setting is important as the conversion from a camera feed to a C-PRID search gallery is not straightforward and needs to be evaluated to design reliable applications. Ideally, from a query image and a raw video feed, a FF-PRID model should find whether or not the query appears in each frame. This way, the system can raise an alert as soon as the searched person is encountered in any camera. But in practice, the FF-PRID task is complex and highly prone to errors. Because of the criticality of the task in many scenarios, the outputs of the model must be cross-checked by a human operator before triggering any action involving security agents.

Thus, we propose an alternative hybrid framework, which requires validation by a human operator after automatic predictions are made by an artificial intelligence model, to address this problem and evaluate it. The proposed pipeline goes as follows: First, the live video stream is cut into short video segments of τ frames. Then, each of these segments are processed by a pedestrian detection model to extract bounding boxes of all the persons present in the video and create a traditional search gallery. The query and the gallery are then processed by a classic Re-ID model and, if the highest similarity score in the gallery is higher than a given threshold β, the η members of the gallery with highest similarity scores are shown to the monitoring agent, who decides if the predictions are correct triggers actions when necessary. The proposed pipeline is illustrated in Fig. 3a. The threshold for raising an alert β, the number of images shown to the agent η and the length of the

video segments τ are user defined parameters that influence the final results. We note that the ideal scenario described above can be obtained with this framework if $\tau = 1$, $\eta = 1$, the FF-PRID works perfectly and β is tuned appropriately.

Validation measures. In the case of a perfect FF-PRID model, the operator validation is required in all the cases where the query is present in the τ frames of video sequence and not in any other case. Hence, there are two ways for a model to fail: by missing the query when it is present in the video segment or by calling the operator when the query is not present. Thus, to evaluate the quality of a model, we define two important indicators that we call *Finding Rate* (FR) and *True Validation Rate* (TVR). They respectively represent the number of sequences in which the query was found when it appeared and the number of times that the query was present when the operator was solicited.

To define these two validation measures formally, some other variables must be introduced first. These variables are influenced by the variables to evaluate the classification task as True Positive (TP), False Negative(FN), True Negative(TN) and False Positive(FP). For a given {query, video} pair, we define:

- **A True Call (TC)**, when the query is present in the video, the highest similarity score is greater than the threshold β and the query is in the top η best candidates. It corresponds to a successful case of re-identification by the system.
- **A True Missed Call (TMC)**, when the query is present in the video, the highest similarity score is greater than β and the query is not in the top η best candidates. It is the case where the query is present, the system is asking for confirmation but does not provide the correct images to the operator and the query is missed anyways.
- **A False Silence (FS)**, when the query is present in the video, but the highest similarity score is smaller than β. It is the case where the query is missed but the operator is not disturbed.
- **A False Call (FC)**, when the query is not in the video but the highest similarity score is greater than β. It corresponds to the case where the operator is disturbed for nothing.
- **A True Silence (TS)**, when the query is not in the video and the highest similarity score is smaller than β. It is the case where the query is not present and nothing happens.

Then, the FR and TVR can be defined as follows:

$$FR = \frac{TC}{TC + TMC + FS}, \tag{1}$$

$$TVR = \frac{TC}{TC + TMC + FC}. \tag{2}$$

FR and TVR are comprised between 0 and 1. Hence, FR = 1 means that whenever the query was present in the video, it was successfully identified by the system (model + operator). Likewise, TVR = 1 means that the operator was never called for nothing, i.e. all the time the model asked for verification, the query was actually present in the proposed cropped images. In contrast, FR < 1 means that in some sequences the query was present but it was missed, and TVR < 1 means that in some situations the model asked for operator validation when the query was not present in the suggestions.

III. Experimental Setup: Dataset and FF-PRID pipeline details

A. Dataset used for validation

To test the proposed framework and metrics, we use a modified version of the PRID-2011 dataset [7], considering raw full frame videos as input instead of the pre-cropped images of the original dataset.

The original PRID-2011 dataset is composed of images extracted from multiple person trajectories recorded from two different static surveillance cameras, named A and B. Images from these cameras contain a view point change and a stark difference in illumination and background. Since images are extracted from trajectories, several successive poses per person are available in each camera view, with some people appearing in both views. After filtering out manually some heavily occluded persons, corrupted images induced by tracking and annotation errors, the official PRID-2011 dataset contains 385 persons in camera view A and 749 in camera view B. The persons with the first 200 labels appear in both views.

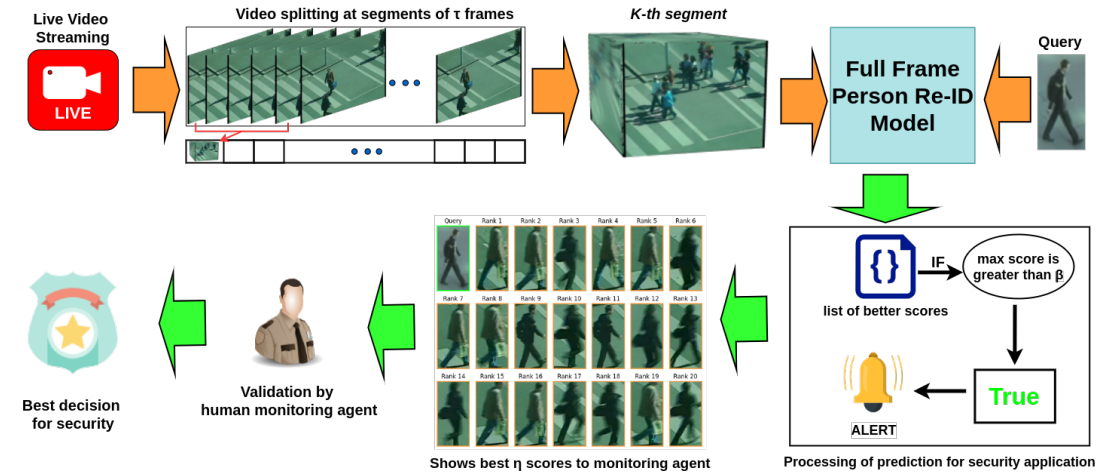
PRID-2011 was created to test classic person Re-ID approaches, as well as video-based Re-ID [9]. To conduct our experiments, we obtained the raw videos and annotations that were used to create the PRID-2011 dataset[2]. From now on, the two raw full frame videos will be called view A (1:01:53 hours) and view B (1:06:39 hours). Both views were cut into sub-videos of 2 minutes, to serve as input to the FF-PRID framework (Fig. 3a). This way, view A contains 30 videos and view B, 33. For each 2 minute video sample, a ground truth file is generated[3].

B. Overview of the Full Frame Re-ID pipeline

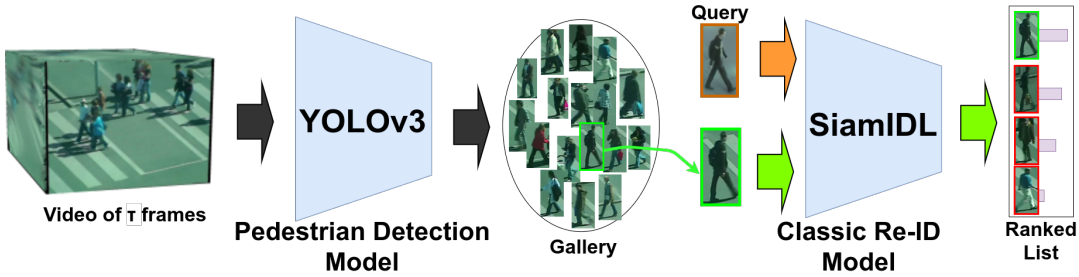
Fig. 3 illustrates the proposed FF-PRID approach. In Fig. 3a, we show the application level Re-ID scheme. The original video is split into shorter sequences and passed to a FF-PRID model. When the model returns a high confidence score that the query is present in the sequence, an alert is raised and a group of persons' images are presented to a monitoring agent for human validation. In Fig. 3b, the FF-PRID model is shown in details. The video is fed to an object detection model in order to detect pedestrians and generate clippings for the search gallery. After this step, the image of the query person is searched in the gallery by means of a classic Re-ID model, which outputs a list of images similar to the query, ordered from most to least similar. Both the pedestrian detection

[2]We kindly thank the authors of the original PRID-2011 paper for their responsiveness and cooperation.

[3]Our scripts for processing the raw videos and generating the ground truth files, as well as the implementation of our baseline pipeline, are openly available at: https://github.com/fsumari/FF-PRID-2020.



(a) Hybrid Human-Machine FF-PRID Framework.



(b) Proposed FF-PRID model

Fig. 3. Hybrid Human-Machine Framework and proposed Pipeline for Full Frame Person Re-Identification.(Source: author)

model and the classic Re-ID model were implemented using TensorFlow 1.14.0 and were executed on a NVIDIA P5000 GPU. We present the implementation of these models in the following subsections.

C. Object Detection

For this work, we use the You Only Look Once (YOLO-v3) [24] approach for pedestrian detection. In short, YOLO methods belong to the family of regression/classification based approaches, mapping directly from image pixels to bounding box coordinates and class probabilities to reduce significantly the time complexity. A detailed explanation of YOLO is out of the scope of this sub-section, and for a complete overview of the recent literature about Object Detection (OD), we refer the reader to the two following surveys [27], [28]. In practice, we use the Darknet-53 architecture and pretrained weights proposed in tensorflow. This network uses 53 convolutional layers with $3x3$ kernels in the beginning and $1x1$ in the end. The model used was trained on the VOC dataset [29], containing 80 classes. Darknet-53 operates at a level close to state-of-the-art object detectors, but is faster because it uses less floating-point operations. The YOLO-v3 model was prepared with a threshold of 0.5 for both Intersection over union (IOU) and the loss function. During our evaluation (Section IV-A), the score threshold to keep a bounding box, as well as the IOU threshold were both set to 0.5 as well. To generate the search galleries, we only use the output corresponding to the person class from the object detector.

D. Classic Person Re-ID

The method used to perform classic person Re-ID in this paper is the same as proposed by [11], called *An Improved Deep Learning Architecture for Person Re-Identification*. From now on, we refer to this method as SiamIDL. This method used the following deep neural network architecture: two layers of tied convolution with max pooling, cross-input neighborhood differences, patch summary features, across-patch features, higher-order relationships, and finally, a softmax function to yield the final estimate of whether the input images are of the same person or not.

For implementation, we used the authors' source code and trained the network using the training set of the CUHK-03 dataset [5]. We use the same parameters as in the original paper: batch_size=50, max_steps=210 000, and learning_rate=0.01. The Cumulative Matching Characteristics (CMC) are computed on both the validation folder of CUHK-03 (938 images) and the original PRID dataset to evaluate the model. Results are presented in Section IV-B. We save final weights to use them to compute over the validation folder. We didn't re-train the model for this step.

IV. Results

To demonstrate the importance of considering the FF-PRID pipeline as a whole, and thus corroborate the usefulness of the proposed metrics, the evaluation conducted in this work is three-fold.

A. Evaluation of the Object Detection model

The PRID-2011 dataset was initially created to evaluate classic Re-ID models. Hence, occluded persons, persons with less than five confidence frames, as well as distorted images caused by tracking and annotation errors were removed from the list of bounding boxes. To achieve a correct evaluation of YOLO-v3 on the PRID-2011 videos, it is necessary to manually add the bounding boxes of these people who were ignored during dataset creation. To do this, the *LabelIMG* tool was used and we added a total of 37.772 bounding boxes for the labels of video B. The results obtained for pedestrian detection with YOLO-v3 on the PRID-2011 videos are presented in Table I. These results correspond to the model that was used to generate the search gallery for the classic Re-ID model(see 3b).

TABLE I
EVALUATION OF THE YOLO-V3 MODEL FOR PEDESTRIAN DETECTION ON THE RAW VIDEO B FROM THE PRID-2011 DATASET. FOR THE ORIGINAL BOUNDING BOXES (OBB) ROWS, METRICS WERE COMPUTED USING ONLY THE BOUNDING BOXES AVAILABLE FROM THE ORIGINAL DATASET AS GROUND TRUTH. FOR THE OBB + MANUALLY ADDED BOUNDING BOXES (MBB) ROWS, THE BOUNDING BOXES ADDED USING THE LABELIMG TOOL WERE ALSO CONSIDERED.

	Precision	**Recall**	**F1-score**	**mAP**
OBB	0.462	0.866	0.603	45.53%
OBB + MBB	0.761	0.824	0.791	69.50%

When analyzing visually the output produced by YOLO-v3, the results on PRID-2011 video B seem almost perfect. In this way, the difference in the results between OBB and OBB+MBB can be interpreted as the number of entire human bodies which where manually filtered by the annotators of the original dataset (e.g. partially overlapping persons). On the other hand, the remaining errors for the OBB+MBB case mostly correspond to incomplete body parts, such as legs, arms or torso, which we did not include in our ground truth bounding boxes (see Fig. 4).

An object detector, such as YOLO-v3, is trained to find the particular characteristics of the object of interest in an image and thus generates bounding boxes for the cases mentioned above. These cases constitute an important discrepancy between the domain on which the classic Re-ID model was trained and the images generated by the OD model. Such domain shift in the inputs of the C-PRID model can be a major source of errors for the full FF-PRID pipeline.

B. Evaluation of the Person Re-ID model

To evaluate the SiamIDL model used in our pipeline, we compute the CMC curves for both the validation set of CUHK-03 and PRID-2011. The evaluation on CUHK-03 was used to validate the training of our model by comparing our results

Fig. 4. Different possible mistakes for cropping. (Source: author).

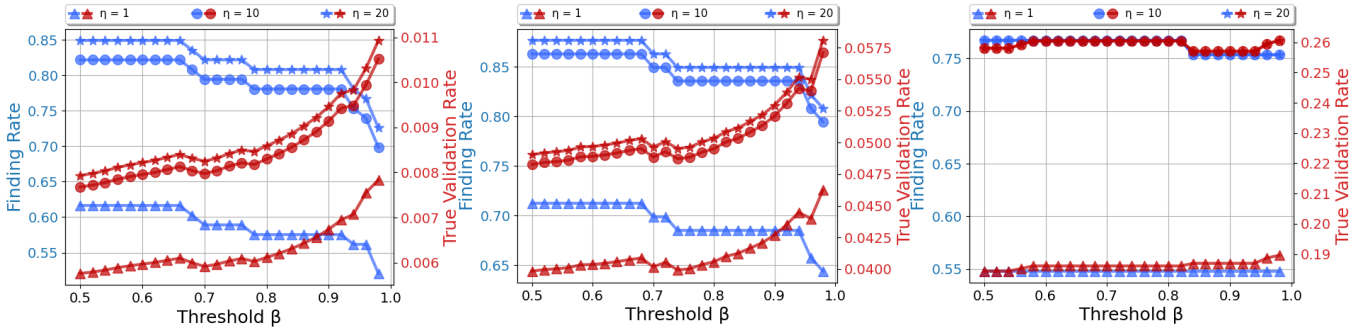
with the ones obtained in the original paper. On the other hand, we test performed on the first 200 Ids from view A of PRID-2011. We were obtained good results with more than 48% on Rank 1 and more than 95% on Rank 20. We note that no additional training was conducted on the PRID-2011 dataset and only the weights trained on CUHK-03 are used in this validation. This last experiment corresponds to the practical scenario of deploying Re-ID in new environments (e.g. new city, new shopping center), where it would be impractical to create a new custom training dataset for every new implementation.

The fact that a network trained on CUHK-03 can generalize to data from another dataset shows that the proposed Re-ID model is able to learn cross-domain Re-ID. This property is interesting as the domains encountered for every new implementation vary a lot depending on the quality of the cameras, the distance to the people and the illumination, among other factors.

C. Evaluation of the full pipeline for FF-PRID

Our evaluation consists of 20 videos and 73 queries (36 for view A and 37 for view B). Each query appears in its associated video at least in one frame, but does not necessarily appear in each sub-videos after splitting into shorter sequences (see Fig. 2). To evaluate the influence of the different parameters of the FF-PRID pipeline, i.e. the number of frames for video splitting τ, the threshold for alert generation β and the number of candidates shown to the monitoring agent η, we use different values for each parameter. Thus, we test $\tau \in \{10, 100, 1000\}$, $\eta \in \{1, 10, 20\}$ and the threshold β is computed for various values in the interval $[0.5, 0.98]$. The Figures. 5a, 5b and 5c shows the Finding Rate (FR) and True Validation Rate (TVR) curves for different values of τ, β and η.

Influence of the FF-PRID parameters. As we can see in these graphs, for all values of τ and η, the FR curves decrease when β increases. This behavior can be explained by the fact that a larger β means that the model will raise less alerts and is more likely to miss the query. However, with $\tau = 1000$, the decreasing effect is less noticeable. This is because when considering larger galleries, the model has more chances of finding a similar image and having at least one high confidence prediction. In contrast, the three TVR curves



(a) Graphic with τ=10. (b) Graphic with τ=100. (c) Graphic with τ=1000.

Fig. 5. Finding Rate (FR) and True Validation Rate (TVR) curves for different values of τ, β and η. *In Fig. 5c, the curves* $\eta = 10$ *and* $\eta = 20$ *are overlapping*.

demonstrate the opposite behavior and are increasing with β. This also makes sense as increasing β correspond to reducing the accepted confidence range and thus calling the agent with less frequency. However, except for the case $\tau = 1000$, we note that the values of the different TVR are all very low, meaning that the human monitoring agent would be called in many unnecessary cases.

Furthermore, as expected, $\eta = 10$ and $\eta = 20$ performed much better than $\eta = 1$ for all configurations of τ and β. Indeed, the C-PRID models are not perfect and training Re-ID models with very high top 1 accuracy is hard. In contrast, decreasing η, reduces the amount of work for the monitoring agent as it needs to control less image samples.

Finally, the FR curves present better results for $\tau = 100$ than for the two other tested values. This is because the raw video is split into sub-videos which are neither too short nor too long. This way, the query appears on the video for a sufficient amount of time to be recognize and there are not too many distractors to confuse the network.

Further considerations. The results obtained for the FF-PRID problem suggest that careful selection of the tunable parameters (τ, β and η) is paramount. Indeed, with proper selection we can reach an FR of almost 80% with a TVR of 26%. Although the score that we managed to reach for the Finding Rate are satisfactory, we acknowledge that the TVR is still too low for the method to be used practically, as the operator would be called too many times if dealing with several cameras at the same time. These mixed results emphasize the importance of considering the FF-PRID problem as a whole and suggest that changing the paradigm for person Re-ID might be the best way to obtain applicable solution for tomorrow's cities.

V. Conclusion

In this work we claim that the classic approach for person Re-ID is not sufficient to develop practical implementations of Re-ID for security application, which requires to process the full frames of the cameras stream (FF-PRID) instead of pre-cropped clean images of people. To support this claim, we build a two steps FF-PRID pipeline. First, persons bounding boxes are extracted from the input video using a state of the art object detection model (YOLO-v3) to generate a search gallery. Then, the query is searched in the gallery using a good Re-ID model (SiamIDL). A framework embedding these two sub-modules is presented, including a human monitoring agent in the loop in order to strengthen the results. We present two new metrics in order to evaluate the proposed FF-PRID pipeline. The metrics are used to evaluate how many times the query is found when it is present in the video (Finding Rate) and how many times the query is present when the agent is solicited (True Validation Rate). These framework and metrics are, to the best of our knowledge, the first proposed approaches to evaluate a FF-PRID model, looking for persons directly in the entire video frames.

Our experimental results were conducted on a modified version of the PRID-2011 dataset. We demonstrated that both the OD model and the classic Re-ID model managed to perform well on the new dataset without additional training. However, the final results for FF-PRID, evaluated using FR and TVR, were not sufficient to deploy FF-PRID in production. Although choosing the right parameters in our framework enabled us to reach a good FR score ($> 80\%$), we were not able to obtain a TVR much better than 25%, which means that most of the time the operator calls were unnecessary. Some possible explanations for these results were discussed as well as possible improvements. However, these mixed results emphasize the importance of considering Re-ID in the FF-PRID setting if we want to develop methods that can be used in practical scenarios. We believe that many improvements could be achieved if the community starts investigating Re-ID solutions for the Full Frame setting instead of focusing only on the classic pre-cropped image-based setting.

A. Future works

A possible direction to achieve this is to consider video-based classic Re-ID methods [9], [10]. Another natural option is to consider the open-world person Re-ID setting instead of closed-world [30]. We also plan to train more specific pedestrian detection techniques, focusing on recognizing only full-bodies.

ACKNOWLEDGMENT

Our work have a published in Pattern Recognition Letters [31], we thank everyone who was part of this research.

REFERENCES

[1] A. Hampapur, L. Brown, J. Connell, S. Pankanti, A. Senior, and Y. Tian, "Smart surveillance: applications, technologies and implications," in *Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint*, vol. 2. IEEE, 2003, pp. 1133–1138.

[2] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *arXiv preprint arXiv:2001.04193*, 2020.

[3] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1116–1124.

[4] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3754–3762.

[5] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 152–159.

[6] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *European conference on computer vision*. Springer, 2008, pp. 262–275.

[7] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Scandinavian conference on Image analysis*. Springer, 2011, pp. 91–102.

[8] A. Brunetti, D. Buongiorno, G. F. Trotta, and V. Bevilacqua, "Computer vision and deep learning techniques for pedestrian detection and tracking: A survey," *Neurocomputing*, vol. 300, pp. 17–33, 2018.

[9] D. Ouyang, Y. Zhang, and J. Shao, "Video-based person re-identification via spatio-temporal attentional and two-stream fusion convolutional networks," *Pattern Recognition Letters*, vol. 117, pp. 153–160, 2019.

[10] S. Li, S. Bak, P. Carr, and X. Wang, "Diversity regularized spatiotemporal attention for video-based person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 369–378.

[11] A. Ejaz, M. Jones, and T. K. Marks, "An Improved Deep Learning Architecture for Person Re-Identification," *Cvpr*, pp. 3908–3916, 2015. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2015/papers/Ahmed_An_Improved_Deep_2015_CVPR_paper.pdf

[12] H. Wang, Y. Fan, Z. Wang, L. Jiao, and B. Schiele, "Parameter-Free Spatial Attention Network for Person Re-Identification," 2018. [Online]. Available: http://arxiv.org/abs/1811.12150

[13] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint Discriminative and Generative Learning for Person Re-identification," 2019. [Online]. Available: http://arxiv.org/abs/1904.07223

[14] Y. Yan, B. Ni, J. Liu, and X. Yang, "Multi-level attention model for person re-identification," *Pattern Recognition Letters*, vol. 127, pp. 156–164, 2019.

[15] N. Martinel, C. Micheloni, and G. L. Foresti, "A pool of multiple person re-identification experts," *Pattern Recognition Letters*, vol. 71, pp. 23–30, 2016.

[16] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016.

[17] R. Satta, F. Pala, G. Fumera, and F. Roli, "Real-time appearance-based person re-identification over multiple kinecttm cameras." in *VISAPP (2)*, 2013, pp. 407–410.

[18] C.-Y. Wang, P.-Y. Chen, M.-C. Chen, J.-W. Hsieh, and H.-Y. M. Liao, "Real-time video-based person re-identification surveillance with light-weight deep convolutional networks," in *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2019, pp. 1–8.

[19] Y. Li, Z. Wu, S. Karanam, and R. J. Radke, "Real-world re-identification in an airport camera network," in *Proceedings of the International Conference on Distributed Smart Cameras*, 2014, pp. 1–6.

[20] O. Camps, M. Gou, T. Hebble, S. Karanam, O. Lehmann, Y. Li, R. J. Radke, Z. Wu, and F. Xiong, "From the lab to the real world: Re-identification in an airport camera network," *IEEE transactions on circuits and systems for video technology*, vol. 27, no. 3, pp. 540–553, 2016.

[21] A. Shenoi, M. Patel, J. Gwak, P. Goebel, A. Sadeghian, H. Rezatofighi, R. Martin-Martin, and S. Savarese, "Jrmot: A real-time 3d multi-object tracker and a new large-scale dataset," *arXiv preprint arXiv:2002.08397*, 2020.

[22] E. Togootogtokh, C. Micheloni, G. L. Foresti, and N. Martinel, "An efficient uav-based artificial intelligence framework for real-time visual tasks," *arXiv preprint arXiv:2004.06154*, 2020.

[23] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[24] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv*, 2018.

[25] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1334–1373, 2016.

[26] R. He, T. Tan, L. Davis, and Z. Sun, "Learning structured ordinal measures for video based face recognition," *Pattern Recognition*, vol. 75, pp. 4–14, 2018.

[27] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *International journal of computer vision*, vol. 128, no. 2, pp. 261–318, 2020.

[28] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

[29] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[30] Q. Leng, M. Ye, and Q. Tian, "A survey of open-world person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[31] F. O. Sumari, L. Machaca, J. Huaman, E. W. Clua, and J. Guérin, "Towards practical implementations of person re-identification from full video frames," *Pattern Recognition Letters*, vol. 138, pp. 513–519, 2020.