Transport and Communications Science Journal

A UNIFIED FRAMEWORK FOR AUTOMATED PERSON RE-IDENTIFICATION

Hong Quan Nguyen[1,3], Thuy Binh Nguyen [1,4], Duc Long Tran [2], Thi Lan Le[1,2]

[1]School of Electronics and Telecommunications, Hanoi University of Science and Technology, Hanoi, Vietnam

[2]International Research Institute MICA, Hanoi University of Science and Technology, Hanoi, Vietnam

[3]Viet-Hung Industry University, Hanoi, Vietnam

[4]Faculty of Electrical-Electronic Engineering, University of Transport and Communications, Hanoi, VietNam

ARTICLE INFO

TYPE: Research Article
Received: 31/8/2020
Revised: 26/9/2020
Accepted: 28/9/2020
Published online: 30/9/2020
https://doi.org/10.47869/tcsj.71.7.11

**Coresponding author:*
Email:thuybinh_ktdt@utc.edu.vn

Abstract. Along with the strong development of camera networks, a video analysis system has been become more and more popular and has been applied in various practical applications. In this paper, we focus on person re-identification (person ReID) task that is a crucial step of video analysis systems. The purpose of person ReID is to associate multiple images of a given person when moving in a non-overlapping camera network. Many efforts have been made to person ReID. However, most of studies on person ReID only deal with well-alignment bounding boxes which are detected manually and considered as the perfect inputs for person ReID. In fact, when building a fully automated person ReID system the quality of the two previous steps that are person detection and tracking may have a strong effect on the person ReID performance. The contribution of this paper are two-folds. First, a unified framework for person ReID based on deep learning models is proposed. In this framework, the coupling of a deep neural network for person detection and a deep-learning-based tracking method is used. Besides, features extracted from an improved ResNet architecture are proposed for person representation to achieve a higher ReID accuracy. Second, our self-built dataset is introduced and employed for evaluation of all three steps in the fully automated person ReID framework.

Keywords. Person re-identification, human detection, tracking

1. INTRODUCTION

Along with the strong development of camera networks, a video analysis system has been become more and more popular and is applied in various practical applications. In early years, these systems are performed in manual manner which are time consuming and tedious. Moreover, the accuracy is low and it is difficult to retrieve information when needed. Fortunately, with the great help of image processing and pattern recognition, automatic techniques are used to solve this problem. The automatic video analysis system normally includes four main components that are object detection, tracking, person ReID, and event/activity recognition. Nowadays, these systems are deployed in airport, shopping mall and traffic management departments [1].

In this paper, we focus on a fully automatic person ReID system which contains only three first steps of the full video analysis system that are person detection, tracking and re-identification. The purpose of human detection is to create a bounding box contain an object in a given image while tracking methods aim at connecting the detected bounding boxes of the same person. Finally, person ReID is to associate multiple images of the same person in different camera views. Although studying on person ReID has achieved some important milestones [2], this problem still has to cope with various challenges, such as the variations in illuminations, poses, view-points, etc.

Additionally, most studies on person ReID only deal with Region of Interests (RoIs) which are extracted manually with high quality and well-alignment bounding boxes. Meanwhile, there are several challenges when working with a unified framework for person ReID in which these bounding boxes are automatically detected and tracked. For example, in the detection step, a bounding box might contain only several parts of the human body, occlusion appears with high frequency, or there are more than one person in a detected bounding box. For the tracking step, the sudden appearance or disappearance of the pedestrian cause the fragment of tracklets and identity switch (ID switch). This makes a pedestrian's tracklet is broken into several fragments or a tracklet includes more than one individual. These errors reduce person ReID accuracy. This is the motivation for us to conduct this study on the fully automated person ReID framework. The contribution of this paper are two-folds. First, a unified framework for person ReID based on deep learning models is proposed. In this framework, among different models proposed for object detection and tracking, YOLOv3 (You Only Look Once) [3] and Mask R-CNN (Mask Region-based Convolution Neural Network) [4] are employed at detection step while DeepSORT [5] is used for tracking thanks to its superior performance [6]. Concerning person re-identification step, features extracted from an improved ResNet architecture are proposed for person representation to achieve a higher ReID accuracy. Second, to evaluate the performance of the proposed system, a dataset is collected and carefully annotated. The performance of the whole system is fully evaluated in this work.

The rest of this paper is organized as follows. Section 2 presents some prominent studies related to a fully automated person ReID system. The proposed framework is introduced in Section 3. Next, several extensive evaluations on each step as well as on overall performance is shown in Section 4. The last Section provides conclusion and future work.

2. RELATED WORK

In this section, some remarkable researches focusing on building a fully automated person ReID system are discussed briefly. First of all, we mention to a study of Pham et al. [7] in which a framework for a fully automated person ReID system including two phases of human detection and ReID is proposed. In this work, in order to improve the performance of human detection an effective shadow removal method based on score fusion of density matching is employed. By this way, the quality of the detected bounding boxes is higher and help to achieve better results on person ReID step. For person ReID step, an improved version of Kernel DEScriptor (KDES) is employ for person representation. Some extensive experiments are conducted on several benchmark datasets and their own dataset to show the effectiveness of the proposed method. In [8], the authors also declare that the quality of human detection impact on person ReID performance. According to the authors, the detected bounding boxes may contain false positive, partially occluded people or are misaligned to the people. In order to tackle the above issues, the authors proposed modifications to classical person detection and re-identification algorithms. However, techniques used in this study are out of date. A unified framework is proposed to tackle both person ReID and camera network topology inference problems in the study of Cho et al. [9]. The initial camera network topology is inferred relied on the obtained results in person ReID task. And then, this initial topology is employed to improve the person ReID performance. This procedure is repeated many time until the estimated camera network topology converges. Once the reliable camera network topology is estimated in the training stage and it can be used for online person ReID and update the camera network topology over time. The proposed framework not only improves person ReID accuracy but also ensures computation speed. However, this work does not mention to two crucial steps in a fully automated person ReID including human detection and tracking. One more work related to the automated person ReID framework we would like to discuss is presented in the PhD thesis of Figueira [10]. In this thesis, the author presents on person ReID problem, its challenges, and existing methods for dealing with this problem. The integration between human detection and person ReID is also examined in this thesis. Nevertheless, the tracking stage and its impact on the person ReID performance is not surveyed.

From the above analysis, we realize that there are a few studies which focus on integrating three main step in the unified person ReID framework. Meanwhile, this is really necessary when building a fully automated person ReID system. This is the motivation for us to perform this research with the two contributions: (1) propose a unified person ReID framework based on deep-learning methods, and (2) introduce our self-built dataset.

3. PROPOSED FRAMEWORK

Figure 1 shows the unified framework for person ReID which includes three crucial steps: human detection, tracking, and person ReID. The purpose of this framework is to evaluate overall performance when all steps are performed in the automatic manner. For human detection step, the two state-of-the-art human detection techniques are proposed to use, that are YOLOv3 [11] and Mask R-CNN [12]. Besides, DeepSORT [5] is adopted in tracking step. Additionally, in order to overcome the challenges caused by human detection and tracking steps, one of the most effective deep-learning features, ResNet is employed for person representation. The effectiveness of ResNet is proved in some existing works [13, 14] In the following sections, we describe briefly the person detection and tracking methods.

3.1. Human detection

In recent years, with the great development of deep-learning networks and the help of computer which has strong computation capability, object detection techniques have achieved high accuracy with real-time response. In the literature, object detection methods are categorized into two main groups: (1) based on classification and (2) based on regression. In the first group, Regions of Interest (RoIs) are chosen and then, they are classified through the help of Convolution Neural Network (CNN). By this way, these methods have to predict which class each selected region belongs to. This leads to time consuming and slow down the detection process. We can list here several methods belonging to this group, such as Region-based Convolutional Neural Network (R-CNN), Fast-RCNN, Faster R-CNN, and Mask R-CNN. In the second group, object detection methods predict bounding boxes and classes in one run of the algorithm. The two most famous techniques belonging to this

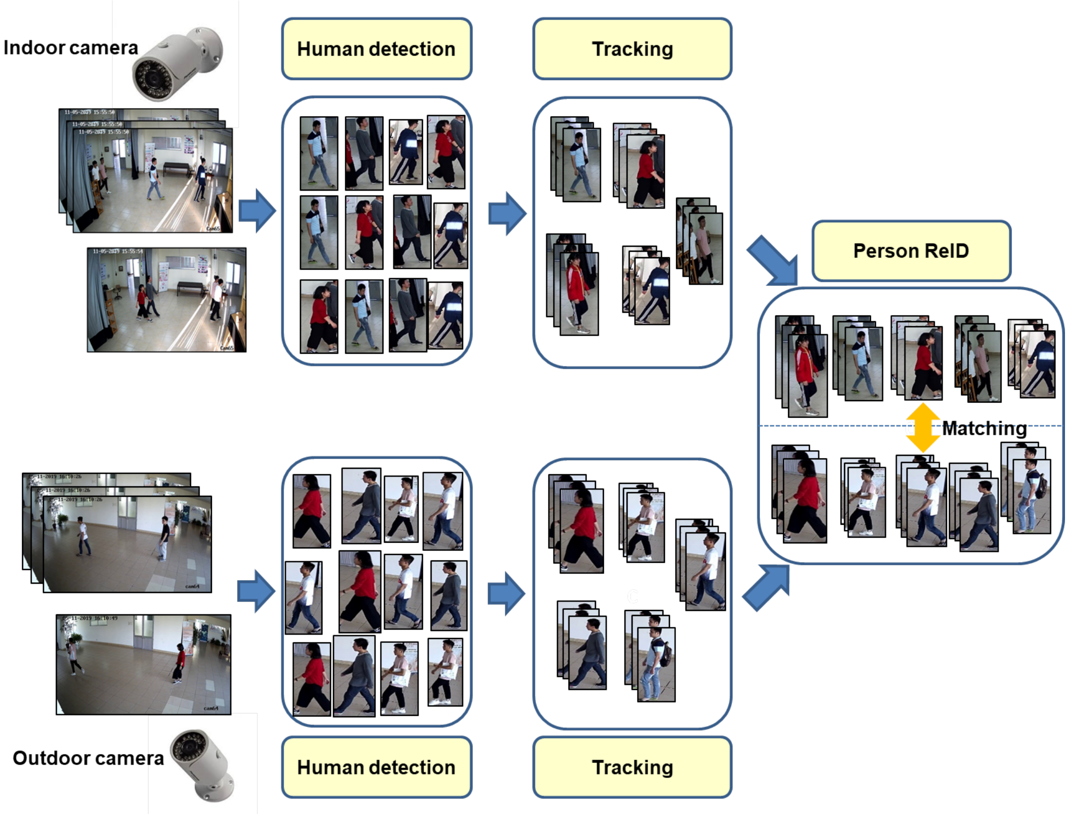


Figure 1. The proposed framework for a fully automated person re-identification.

group are YOLO [11] and Single Shot Multibox Detector (SSD). Among these techniques, YOLO and Mask R-CNN are proposed to employ in this study because of their advantages.

3.1.1. YOLO

Up to now, YOLO [11] has been developed with three versions including YOLOv1, YOLOv2, and YOLOv3. In YOLO algorithm, each bounding box is represented by four descriptors including center of a bounding box, its width, height and class of an detected object. In comparison with the other versions, YOLOv3 has the highest speed and is able to detect a small-size object thanks to a more complicated structure with Pyramid features.

3.1.2. Mask R-CNN

Mask R-CNN [4] is an improved version of Faster R-CNN [12] with the capability of generating simultaneously a bounding box and its corresponding mask for a detected object. The outstanding of Mask R-CNN is to integrate an object proposal generator into a detection network. This help to share deep-learning features between the object proposals and detection networks which lead to reduce computation cost but increase mean Average Precision (mAP) gain.

3.2. Human tracking

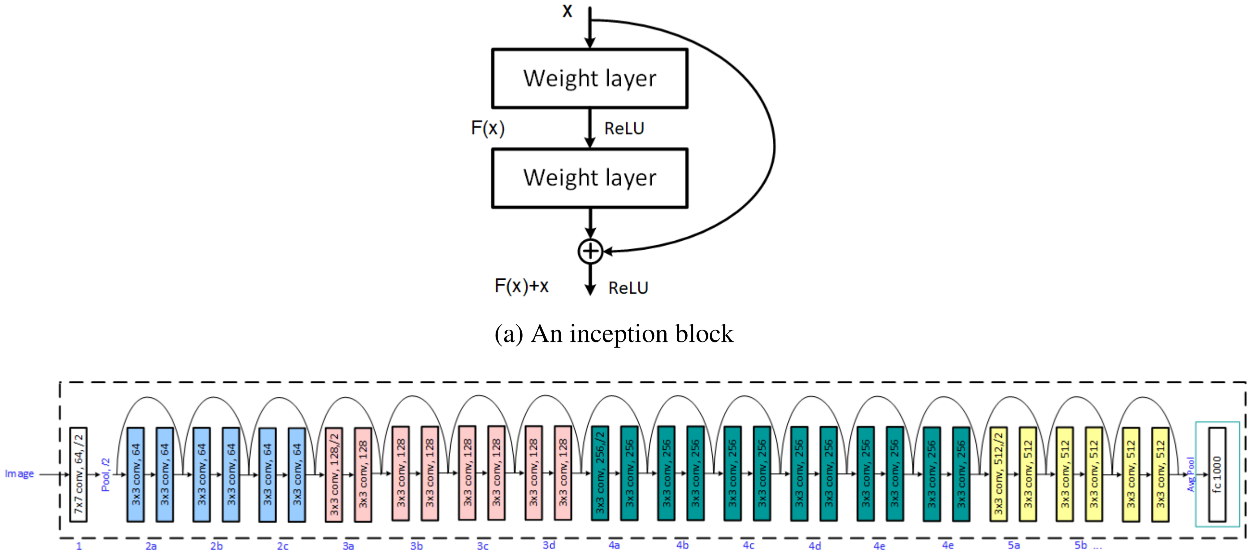
Some earlier works only focus on building a robust and effective detector which needs to scan every frame to find out the regions of interests (ROIs). However, by coupling human detection and tracking together is able to improve the performance of a surveillance system. Instead of scanning every frame, the detector only works on every five frames or moreover. This leads to reduce significantly computation time as well as memory storage. Furthermore, tracking also increases the accuracy of a detector when occlusion appears.

DeepSORT is developed from Simple Online and Realtime Tracking (SORT) [5] which based on Kalman filter [15]. The advantage SORT is to have high speed but ensure high performance. However, a backward of this algorithm is to create a large number of identity switch (IDSW) errors due to occlusion. In order to overcome this issue, DeepSORT extracts appearance features for person representation by adding a deep network which is pre-trained on a large-scale dataset. In DeepSORT algorithm, the distance between the *i*-th track and the *j*-th detected bounding box is defined as shown in Eq. (1):

$$c_{i,j} = \lambda d^{(1)}(i,j) + (1-\lambda) d^{(2)}(i,j), \tag{1}$$

where $d^{(1)}(i,j)$ and $d^{(2)}(i,j)$ are the two distances calculated through motion and appearance information, respectively. While $d^{(1)}(i,j)$ is calculated based on Mahalanobis distance, $d^{(2)}(i,j)$ is the smallest cosine distance between the *i*-th track and the *j*-th detected bounding boxes in the appearance space; hyperparameter λ controls this association.

3.3. Person ReID



(a) An inception block

(b) Overall structure of ResNet-50

Figure 2. Structure of a) an inception block b) ResNet-50 [16].

Person ReID is the last step in a fully automated person ReID system. The performance of this step strongly depends on quality of the two previous steps (human detection and tracking). In the literature, most studies on person ReID have paid attention on either feature extraction or metric learning approach. There are a large number of features are designed for person representation. They are classified into two main categorizes: hand-designed and deep-learned features. Hand-designed features mainly rely on researchers' experience and contain information about color, texture, shape, etc. While deep-learned features based on pre-trained model which is generated from the training phase. In this paper, an improved version ResNet feature is proposed for person representation.

The outstanding point of ResNet is to have a deep structure with multiple stacked layers. However, it is not easy to increase the number of layers in a convolutional neural network due to vanishing gradient problem. Fortunately, with the appearance of skip connections which couple the current layer with the previous layer, as shown in Fig. 2a). With the deep structure, ResNet is proposed to use in different pattern recognition, such as object detection, face recognition, image classification, etc. In our work, ResNet-50 is employed. The architecture of this network is illustrated in Fig.

2b). A given image is divided into seven overlapping regions. These regions are fed into ResNet model that is pretrained on ImageNet dataset. 2048-dimensional vector is extracted from the last Convolution layer of the ResNet architecture. Then, ResNet features extracted on each region are concatenated to form a final feature vector for person representation. By this way, feature vectors take into account the natural relation between between human parts.

4. EXPERIMENTAL RESULTS

4.1. Datasets

To best of our knowledge, there has been no dataset used for evalutating performance of all three steps in the fully automated person ReID framework. Most existing datasets are only utilized in one of three considered tasks. Therefore, in this study, a dataset is built by our own for evaluating the performance of each step in the fully automated person ReID system, called *Fully Automated Person ReID (FAPR)* dataset. This dataset contains total 15 videos and is recorded on three days by two static non-overlapping cameras with HD resolution (1920×1080), at 20 frames per second (fps) in indoor and outdoor environment conditions. Some descriptions about this dataset is shown in Table 1 in different terms: *#Images, #Bounding boxes, #BB/Images, #IDs, #Tracklets*. Some characteristics of this dataset are described as follows.

Firstly, due to the limitation of observation environment, the distances from pedestrians to cameras are not far (about from 2 meters to 8 meters). This leads to strong variation in human body scale in a captured image. Secondly, the border area of each extracted image is blurred because of pedestrian movement and low quality of surveillance cameras. The blurred phenomenon also causes a great difficulty for human detection as well as tracking steps. Thirdly, two cameras are installed to observe pedestrians horizontally. Lastly, as above mentioned, this dataset is captured in both indoor and outdoor environments. The videos captured in indoor are suffered from neon light, while outdoor videos are collected without daylight with heavy shadow. Especially, three videos (*20191105_indoor_left, 20191105_indoor_right, 20191105_indoor_cross*) are captured by sunlight which cause noise for all steps. All characteristics mentioned above make this dataset also contains common challenges as existing datasets used for human detection, tracking and person ReID. In order to generate the ground truth for human detection evaluation, bounding boxes are manually created by LabelImg tool [17] which is the widely used tool for image annotation. Five annotators have prepared all groundtruth for person detection, tracking and re-identification.

Table 1. Sample video descriptions.

Videos	#Images	#Bounding boxes	#BB/Image	#IDs	#Tracklets
indoor	489	1153	2.36	7	7
outdoor_easy	1499	2563	1.71	6	7
outdoor_hard	2702	6552	2.42	8	20
20191104_indoor_left	363	1287	3.55	10	10
20191104_indoor_right	440	1266	2.88	10	13
20191104_indoor_cross	240	1056	4.40	10	10
20191104_outdoor_left	449	1333	2.97	10	10
20191104_outdoor_right	382	1406	3.68	10	11
20191104_outdoor_cross	200	939	4.70	10	12
20191105_indoor_left	947	1502	1.59	10	11
20191105_indoor_right	474	1119	2.36	10	10
20191105_indoor_cross	1447	3087	2.13	10	21
20191105_outdoor_left	765	1565	2.05	11	11
20191105_outdoor_right	470	1119	2.38	10	11
20191105_outdoor_cross	1009	2620	2.60	9	17

4.2. Evaluation measures

In this section, different evaluation measures are employed to show the performance of each step in the fully automated person ReID framework. It is worth noting that evaluation measures for human detection and tracking are described in details [6, 18]. Those measures are also used for evaluating the performance of human detection and tracking in this paper. Concerning person ReID, Cumulative Matching Characteristic (CMC) curve is utilized for person ReID evaluation. These measures are briefly described as follows:

4.2.1. Evaluation measures for human detection

Two measures that are Precision (Prcn) and Recall (Rcll) are used for evaluating human detection. These two metrics are computed in Eq. 2:

$$Prcn = \frac{TP}{TP + FP} \qquad Rcll = \frac{TP}{TP + FN} \tag{2}$$

where, TP, FP, and FN means the number of True Positive, False Positive, and False Negative bounding boxes, respectively. Noted that a bounding box is considered as a TP if it has $\mathrm{IoU} \geq 0.5$ where IoU is the ratio of Intersection over Union between detected bounding box and its corresponding ground-truth.

Besides, F1-score is also used for detection evaluation. This measurement is defined as the harmonic mean of Prcn and Rcll as shown in Eq. (3):

$$F1 - score = 2 \times \frac{Prcn \times Rcll}{Prcn + Rcll} \tag{3}$$

4.2.2. Evaluation measures for human tracking

We employ different measures to evaluate the performance of a human tracking method as follows:

- **IDP (ID Precision) and IDR (ID Recall)**
The two measures have the same meaning to Prcn and Rcll in object detection evaluation. They are defined as in Eq. (4).

$$IDP = \frac{IDTP}{IDTP + IDFP} \quad IDR = \frac{IDTP}{IDTP + IDFN} \tag{4}$$

where, IDTP: sum of TP in detection and the number of correctly labeled objects in the tracking; IDFP/IDFN: sum of FP/FN in detection and the number of correctly predicted objects for positive class in detection but incorrectly labeled in tracking.

- **IDF1**: This metric is formulated based on IDP and IDR as in Eq.(5). The higher IDF1 is, the better tracker is.

$$IDF1 = 2 \times \frac{IDP \times IDR}{IDP + IDR} \tag{5}$$

- **ID switch (IDs)**: The number of identity switches in total tracklets. This metric means that several individuals are assigned to the same label.

- **The number of track fragmentations (FM)**: This value counts how many times a groundtruth trajectory is interrupted.

- **MOTA (Multi Object Tracking Accuracy)**: This is the most important metric for object tracking evaluation. MOTA is defined as:

$$MOTA = 1 - \frac{\sum_t (IDFN_t + IDFP_t + IDs_t)}{\sum_t GT_t}, \tag{6}$$

  where, t is the index of frame, GT is the number of observed objects in the real-world. It is worth to note that MOTA would be a negative value if there are many errors in the tracking process and the number of these errors is larger than that of observed objects.

- **MOTP (Multi Object Tracking Precision)**: MOTP is defined as the average distance between all true positive and their corresponding ground truth targets.

$$MOTP = \frac{\sum_{t,i} d_{t,i}}{\sum_t c_t} \tag{7}$$

  where, c_t denotes the number of matches found in frame t and $d_{t,i}$ is the sum of distances between all true positives and their corresponding ground truth i. This metric indicates the ability of the tracking in estimating precise object positions.

- **Track quality measures**: Beside the above parameters, three metrics that are mostly tracked (MT), partially tracked (PT), mostly loss (ML) tracklets are also used for tracking evaluation. A target is mostly tracked if its tracking time is at least 80% total length of the ground truth trajectory. While, if a track is only covered for less than 20%, it is called mostly lost. The other cases are defined as partially tracked.

4.2.3. Evaluation measures for person re-identification

In order to evaluate the proposed methods for person ReID, we used Cumulative Matching Characteristic (CMC) curves [19]. CMC shows a ranked list of retrieval person based on the similarity between a gallery and a query person. The value of the CMC curve at each rank is the rate of the true matching results and total number of queried persons. The matching rates at several important ranks (1, 5, 10, 20) are usually used for evaluating the effectiveness of a certain method.

4.3. Experimental results and Discussions

In this section, the obtained results on a unified person ReID framework are shown. It is worth noting that all experiments are conducted on a computer with Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz, 6 cores, 12 threads, RAM 32GB, GPU 1080Ti. Our framework based on Keras with backend Tensorflow, Ubutun 18.4, Python 3. Some parameters in our experiments as follows: size of input images is 1920×1080, $\mathrm{sampling} = 2$, $\mathrm{down\_sample\_ratio} = 1$, $\mathrm{IoU\_threshold} = 0.5$. Since the pedestrian's movement speed is not so fast, the difference between two consecutive frames is not significant. Therefore, in the detection step, we chose $\mathrm{sampling} = 2$ to speed up computation processing.

First, we pay attention on human detection and tracking evaluation on FAPR dataset. In order to show the effectiveness of different coupling of human detection and tracking methods, YOLOv3 and Mask R-CNN are proposed to use in the human detection step, while DeepSORT is employed in the tracking step. Noted that YOLOv3 and Mask R-CNN networks are pre-trained on VOC and MS COCO datasets, respectively. Tables 2 and 3 provide some outcomes of the human detection and tracking tasks. For human detection evaluation, we pay more attention on Precision (Prcn) and Recall (Rcll). Higher Prcn or Rcll is achieved with better human detector. Depending on the characteristic of each video, these values are different from each other. By observing the two

Table 2. Performance on FAPR dataset when employing YOLOv3 as a detector and DeepSORT as a tracker.

Videos	For evaluating a detector (1)					For evaluating a tracker (2)										
	FP↓	FN↓	Rcll(%)↑	Prcn(%)↑	F1-score(%)↑	GT	MT↑	PT↑	ML↓	IDF1(%)↑	IDP(%)↑	IDR(%)↑	IDs↓	FM↓	MOTA(%)↑	MOTP↓
indoor	80	51	95.6	93.2	94.4	7	7	0	0	91.5	90.4	92.7	7	11	88.0	0.26
outdoor_easy	70	65	97.5	97.3	97.4	7	7	0	0	74.5	74.4	74.6	6	16	94.5	0.21
outdoor_hard	533	460	93.0	92.0	92.5	20	19	1	0	78.0	77.6	78.4	30	67	84.4	0.28
20191104_indoor_left	164	215	83.3	86.7	85.0	10	8	2	0	83.8	85.5	82.1	7	24	70.0	0.34
20191104_indoor_right	118	188	85.2	90.1	87.6	13	8	5	0	79.6	81.9	77.4	9	16	75.1	0.30
20191104_indoor_cross	142	244	76.9	85.1	80.8	10	5	4	1	68.0	71.6	64.7	12	29	62.3	0.29
20191104_outdoor_left	249	160	88.0	82.5	85.2	10	8	2	0	73.5	71.2	76.0	10	48	68.6	0.33
20191104_outdoor_right	203	197	86.0	85.6	85.8	11	7	3	1	70.6	70.5	70.8	17	45	70.3	0.29
20191104_outdoor_cross	213	134	85.7	79.1	82.3	12	8	2	2	71.9	69.2	75.0	14	33	61.6	0.30
20191105_indoor_left	66	276	81.6	94.9	87.7	11	6	4	1	84.1	90.9	78.2	14	34	76.3	0.29
20191105_indoor_right	106	291	74.0	88.7	80.7	11	5	6	0	77.4	85.1	71.0	7	49	63.9	0.32
20191105_indoor_cross	284	833	73.0	88.8	80.1	21	10	11	0	68.7	76.1	62.6	29	104	62.9	0.28
20191105_outdoor_left	104	104	93.4	93.4	93.4	11	10	1	0	92.1	92.1	92.1	8	24	86.2	0.27
20191105_outdoor_right	220	256	77.1	79.7	78.4	11	4	6	1	67.3	68.4	66.2	14	67	56.2	0.33
20191105_outdoor_cross	317	378	85.6	87.6	86.6	17	15	2	0	72.2	72.8	71.4	48	97	71.6	0.29
OVERALL	**2869**	**3852**	**86.5**	**89.6**	**88.0**	**182**	**127**	**49**	**6**	**76.6**	**77.9**	**75.3**	**232**	**664**	**75.7**	**0.28**

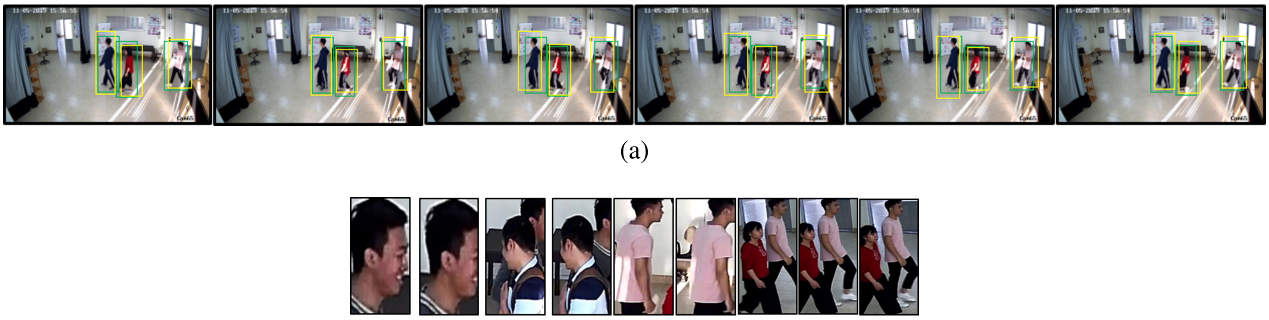
Tables 2 and 3, we realize that Prcn is in range from 79.1% to 97.3% and from 79.9% to 94.4% when applying YOLOv3 and Mask R-CNN, respectively while Rcll varies from 73.0% to 97.5% and from 82.8% to 98.4% in case of using YOLOv3 and Mask R-CNN, respectively. The large difference between these results indicate the great difference in challenging levels of each video.

Table 3. Performance on FAPR dataset when employing Mask R-CNN as a detector and DeepSORT as a tracker.

Videos	For evaluating a detector (1)					For evaluating a tracker (2)										
	FP↓	FN↓	Rcll(%)↑	Prcn(%)↑	F1-score(%)↑	GT	MT↑	PT↑	ML↓	IDF1(%)↑	IDP(%)↑	IDR(%)↑	IDs↓	FM↓	MOTA(%)↑	MOTP↓
indoor	87	18	98.4	92.9	95.6	7	7	0	0	92.7	90.1	95.5	2	6	90.7	0.22
outdoor_easy	148	47	98.2	94.4	96.3	7	7	0	0	93.6	91.8	95.5	2	10	92.3	0.18
outdoor_hard	569	226	96.6	91.7	94.1	20	19	1	0	85.3	83.2	87.5	13	29	87.7	0.26
20191104_indoor_left	128	93	92.8	90.3	91.5	10	9	1	0	91.0	89.8	92.2	5	18	82.4	0.31
20191104_indoor_right	175	46	96.4	87.5	91.7	13	12	1	0	82.8	78.9	87.0	12	14	81.6	0.26
20191104_indoor_cross	165	89	91.6	85.4	88.4	10	9	1	0	72.1	69.7	74.7	15	29	74.5	0.27
20191104_outdoor_left	217	28	97.9	85.7	91.4	10	10	0	0	91.0	85.3	97.4	2	12	81.5	0.28
20191104_outdoor_right	275	169	88.0	81.8	84.8	11	8	2	1	74.5	71.9	77.3	13	33	67.5	0.26
20191104_outdoor_cross	244	75	92.0	78.0	84.4	12	9	3	0	67.6	62.5	73.7	22	20	63.7	0.27
20191105_indoor_left	130	140	90.7	91.3	91.0	11	9	2	0	87.8	88.0	87.5	14	35	81.1	0.27
20191105_indoor_right	143	164	85.3	87.0	86.1	11	8	3	0	80.5	81.2	79.7	7	41	71.9	0.30
20191105_indoor_cross	520	531	82.8	83.1	82.9	21	14	7	0	74.4	74.4	74.2	45	112	64.5	0.27
20191105_outdoor_left	229	37	97.6	87.0	92.0	11	10	1	0	90.1	85.1	95.6	5	8	82.7	0.22
20191105_outdoor_right	240	164	85.3	79.9	82.5	11	6	5	0	73.8	71.4	76.3	12	59	62.8	0.31
20191105_outdoor_cross	370	243	90.7	86.5	88.6	17	17	0	0	75.2	73.2	77.1	37	81	75.2	0.25
OVERALL	**3640**	**2070**	**92.8**	**87.9**	**90.3**	**182**	**154**	**27**	**1**	**82.8**	**80.6**	**85.1**	**206**	**507**	**79.3**	**0.26**

Among 15 considered videos, three videos are most challenging including *20191105_indoor_right, 20191105_indoor_cross and 20191105_outdoor_right*. The most tracked tracklets for those videos are 45.45%, 47.62%, 36.36% and 72.73%, 66.67%, 54.54% compared to the highest result (100%) when coupling YOLOv3 and Mask R-CNN with DeepSORT, respectively. This is also shown through MOTA and MOTP values. When working on *20191105_outdoor_right*, MOTA and MOTP are 56.2% and 0.33, 62.80% and 0.31 in the two examined cases YOLOv3 and Mask R-CNN, respectively. This can be explained that this video has 10 individuals but there are six persons (three pairs) move together which cause serious occlusions in a long time. Therefore, it is really difficult to detect human regions as well as to track pedestrian's trajectories.

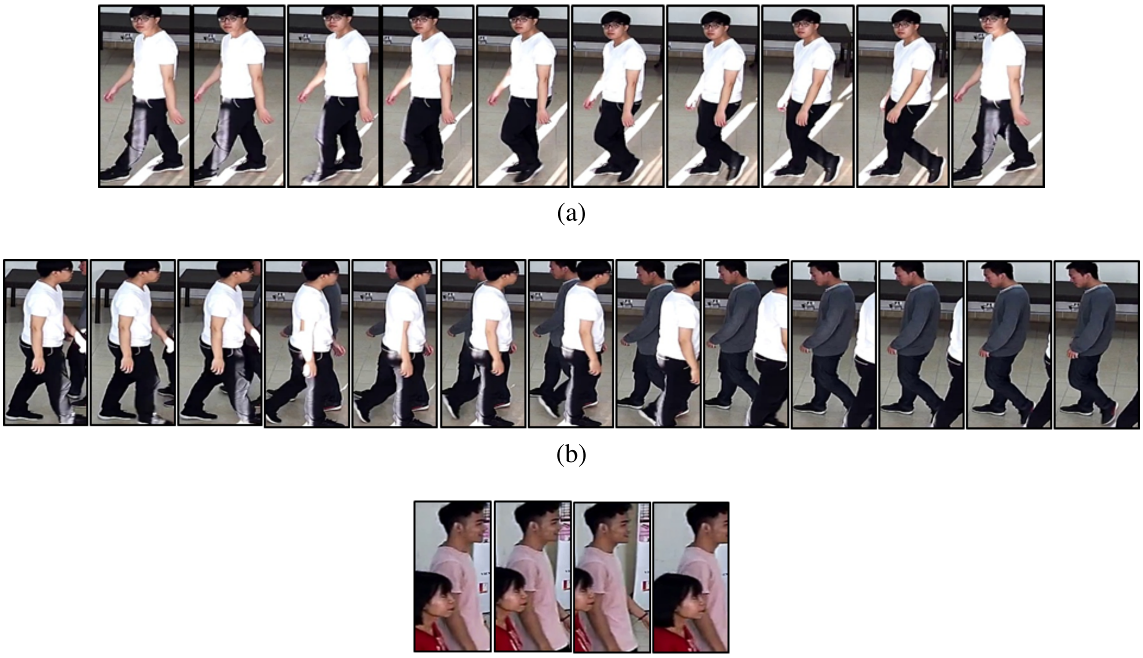
One interesting point is that the best results obtained when working on *outdoor_easy* video, MOTA and MOTP are 94.5% and 0.21, 92.3% and 0.18 in case of applying YOLOv3 or Mask R-CNN for human detection and DeepSORT for tracking, respectively. These values show the effectiveness of the proposed framework for both human detection and tracking steps with high accuracy but small average distance between all true positive and their corresponding target. Figures 3 and 4 show several examples for obtained results in human detection and tracking steps.



(a)

(b)

Figure 3. An example indicates the obtained results in human detection. a) The detected boxes and their corresponding ground-truth are remarked in green and yellow bounding boxes, respectively. b) several errors appeared in human detection step: human body-part detection or a bounding box contains more than one pedestrian.



(a)

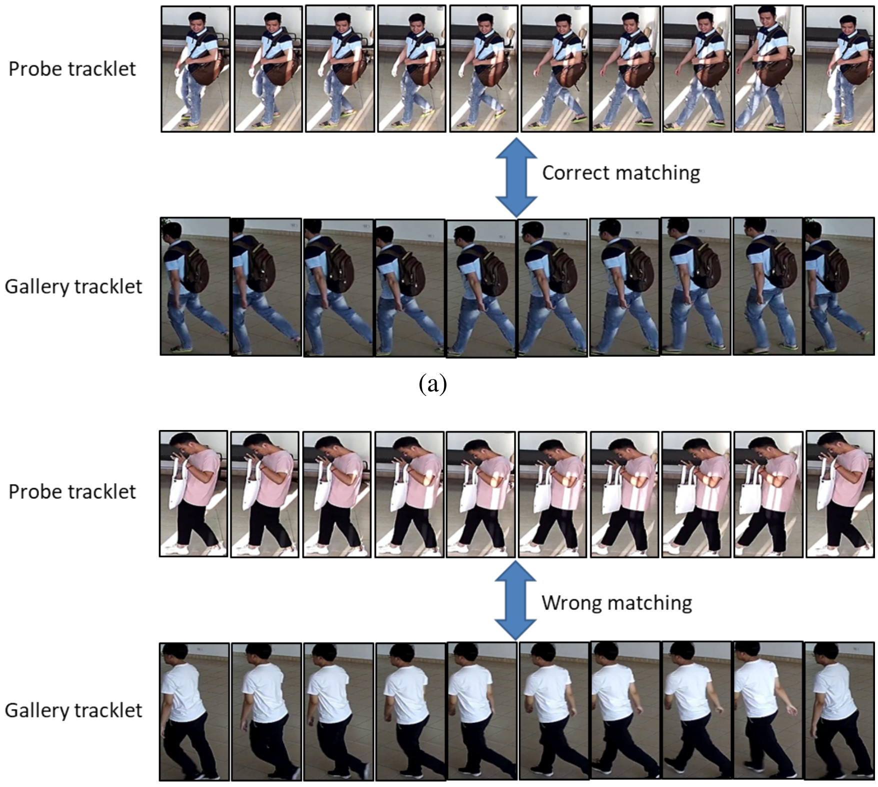
(b)

(c)

Figure 4. An example for obtained results tracking step a) a perfect tracklet, b) switch ID, and c) a tracklet has a few bounding boxes.

Concerning Person ReID, in this study, ResNet features are proposed for person representation and similarities between tracklets are computed based on cosine distance. For feature extraction step, ResNet-50 [20] is pre-train on ImageNet [21], a large-scale and diversity dataset designed for use in visual object recognition research, and then fined tune on PRID-2011 [22] for person ReID task. For tracklet representation, ResNet feature is first extracted on every bounding box belonging to the same tracket. These extracted features are forward to temporal feature pooling layer to generate the final feature vector. For image representation, in order to exploit both local and global information of an image, each image is divided into seven non-overlapping regions. Feature is extracted on each region and then, the extracted features are concatenated together to form a large-dimensional vector for image representation. By this way, we can achieve more useful information and improve the matching rate for person ReID.

For person ReID evaluation, 12 videos (shown in the below Table) are used, in which half of these videos are captured on the same day by the two indoor and outdoor cameras in three different scenarios according to movement manners (left, right, and cross). In the proposed framework for automated person ReID, the matching problem is considered as tracklet matching. For this, indoor tracklets are used as the probe set and outdoor tracklets are the gallery set. The experiments are performed in both cases including single view and multi-views. Due to limitation of research time, in this work, we only focus on matching rate at rank-1. The matched tracklet for the given probe tracklet is chosen based on taking the minimum distance between the probe tracklet and each of gallery tracklets. These matched pairs are divided into correct and wrong matching. A matched pair is called correct matching if these two tracklets represent the same pedestrian. Inversely, if the matched pair describes different pedestrians, it called wrong matching. True and wrong matching are described in Fig. 5. The matching rate at rank-1 is the ratio between the number of correct matching and total of the probe tracklets. The obtained results show that the matching rates for the last four cases (in the same day) are higher than the others and when pedestrians move cross each other, the ReID performance is worst. Additionally, even in case of mixed data (for all movement direction on the same day), the matching rates are 58.82% and 78.57%. These results bring hopefulness for building a fully automated person ReID in practice.



(a)

(b)

Figure 5. An example for obtained results in person ReID step a) true matching and b) wrong matching.

5. CONCLUSIONS

This paper proposes a unified framework for automated person ReID. The contribution of this paper are two-folds. First, deep-learning based methods are proposed for all three steps of this

Table 4. Matching rate (%)at rank-1 for person ReID task in different scenarios.

Scenarios	Probe	Gallery	Matching rates (%)
1	20191104_indoor_left	20191104_outdoor_left	53.33
2	20191104_indoor_right	20191104_outdoor_right	64.29
3	20191104_indoor_cross	20191104_outdoor_cross	45.45
4	20191104_indoor_all	20191104_outdoor_all	58.82
5	20191105_indoor_left	20191105_outdoor_left	100.00
6	20191105_indoor_right	20191105_outdoor_right	75.00
7	20191105_indoor_cross	20191105_outdoor_cross	57.14
8	20191105_indoor_all	20191105_outdoor_all	78.57

framework. The combination of YOLOv3 or Mask R-CNN and DeepSORT for human detection and tracking, respectively. Meanwhile, in person ReID step, the improved version of ResNet features with 7-stripes are used for person representation. Second, FAPR dataset is built on our own for evaluating performance of all three steps. This dataset has the same challenging compared to the common used datasets. The obtained results bring the feasibility of building a fully automated person ReID system in practical. However, the examined videos in this study contain a few persons leading a non-objective results. In the future work, we will consider this issue and deal with complicated data.

ACKNOWLEDGMENT.

This research is funded by University of Transport and Communications (UTC) under grant number T2020-DT-003.

REFERENCES

[1] M. Zabłocki, K. Gościewska, D. Frejlichowski, R. Hofman, Intelligent video surveillance systems for public spaces–a survey, Journal of Theoretical and Applied Computer Science 8 (4) (2014) 13–27.

[2] Q. Leng, M. Ye, Q. Tian, A survey of open-world person re-identification, IEEE Transactions on Circuits and Systems for Video Technology 30 (2019) 1092–1108. https://doi.org/10.1109/TCSVT.2019.2898940.

[3] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, arXiv preprint arXiv:1804.02767, 2018. https://arxiv.org/pdf/1804.02767v1.pdf.

[4] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.

[5] N. Wojke, A. Bewley, D. Paulus, Simple online and realtime tracking with a deep association metric, in: 2017 IEEE International Conference on Image Processing (ICIP), 2017, pp. 3645–3649. https://doi.org/10.1109/ICIP.2017.8296962.

[6] H.-Q. Nguyen, T.-B. Nguyen, T.-A. Le, T.-L. Le, T.-H. Vu, A. Noe, Comparative evaluation of human detection and tracking approaches for online tracking applications, in: 2019 International Conference on Advanced Technologies for Communications (ATC), IEEE, 2019, pp. 348–353. https://www.researchgate.net/publication/336719645 Comparative evaluation of human detection and tracking approaches for online tracking applications.pdf.

[7] T. T. T. Pham, T.-L. Le, H. Vu, T. K. Dao, et al., Fully-automated person re-identification in multi-camera surveillance system with a robust kernel descriptor and effective shadow removal method, Image and Vision Computing 59 (2017) 44–62. https://doi.org/10.1016/j.imavis.2016.10.010.

[8] M. Taiana, D. Figueira, A. Nambiar, J. Nascimento, A. Bernardino, Towards fully automated person re-identification, in: 2014 International Conference on Computer Vision Theory and Applications (VISAPP), Vol. 3, IEEE, 2014, pp. 140–147. https://ieeexplore.ieee.org/document/7295073.

[9] Y.-J. Cho, J.-H. Park, S.-A. Kim, K. Lee, K.-J. Yoon, Unified framework for automated person re-identification and camera network topology inference in camera networks, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 2601–2607. https://arxiv.org/abs/1704.07085.

[10] D. A. B. Figueira, Automatic person re-identification for video surveillance applications, Ph.D. thesis, University of Lisbon, Lisbon, Portugal (2016). https://www.ulisboa.pt/prova-academica/automatic-person-re-identification-video-surveillance-applications.

[11] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788. https://arxiv.org/abs/1506.02640.

[12] S. Ren, K. He, R. Girshick, 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. https://arxiv.org/abs/1506.01497.

[13] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, R. J. Radke, A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets, IEEE Transactions on Pattern Analysis & Machine Intelligence (1) (2018) 1–1.

[14] L. Zheng, Y. Yang, A. G. Hauptmann, Person re-identification: Past, present and future, arXiv preprint arXiv:1610.02984. https://arxiv.org/pdf/1610.02984.pdf.

[15] R. E. Kalman, A new approach to linear filtering and prediction problems, Journal of basic Engineering 82 (1) (1960) 35–45. https://doi.org/10.1109/9780470544334.ch9.

[16] M. ul Hassan, ResNet (34, 50, 101): Residual CNNs for Image Classification Tasks. https://neurohive.io/en/popular-networks/resnet/, [Online; accessed 10-March-2020].

[17] Tzutalin, Labelimg. gitcode(2015). https://github.com/tzutalin/labelImg/, [Online; accessed 20-Sep-2020].

[18] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, K. Schindler, Mot16: A benchmark for multi-object tracking, arXiv preprint arXiv:1603.00831, 2016. https://arxiv.org/abs/1603.00831

[19] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, P. Tu, Shape and appearance context modeling, in: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, IEEE, 2007, pp. 1–8. https://www.ndmrb.ox.ac.uk/research/our-research/publications/439059

[20] J. Gao, R. Nevatia, Revisiting temporal modeling for video-based person ReID, arXiv preprint arXiv:1805.02104. https://arxiv.org/abs/1805.02104

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, IEEE, 2009, pp. 248–255.
https://www.bibsonomy.org/bibtex/252793859f5bcbbd3f7f9e5d083160acf/analyst

[22] M. Hirzer, C. Beleznai, P. M. Roth, H. Bischof, Person re-identification by descriptive and discriminative classification, in: Scandinavian conference on Image analysis (2011), Springer, 2011, pp. 91–102. https://doi.org/10.1007/978-3-642-21227-7_9