Pattern Recognition Letters 143 (2021) 50–57

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

SSS-PR: A short survey of surveys in person re-identification ☆

Ehsan Yaghoubi [a,*], Aruna Kumar [b], Hugo Proença [a]

[a] IT: Instituto de Telecomunicações, University of Beira Interior, Covilhã, Portugal
[b] University of Beira Interior, Covilhã, Portugal

ARTICLE INFO

Article history:
Received 22 May 2020
Revised 2 December 2020
Accepted 23 December 2020
Available online 8 January 2021

MSC:
41A05
41A10
65D05
65D17

Keywords:
Person re-identification
Privacy and security
Visual surveillance

ABSTRACT

Person re-identification (re-id) addresses the problem of whether "*a query image corresponds to an identity in the database*" and is believed to play a fundamental role in security enforcement in the near future, particularly in crowded urban environments. Due to many possibilities in selecting appropriate model architectures, datasets, and settings, the performance reported by the state-of-the-art re-id methods oscillates significantly among the published surveys. Therefore, it is difficult to understand the mainstream trends and emerging research difficulties in person re-id. This paper proposes a multi-dimensional taxonomy to categorize the most relevant researches according to different perspectives and tries to unify the categorization of re-id methods and fill the gap between the recently published surveys. Furthermore, we discuss the open challenges with a focus on privacy concerns and the issues caused by the exponential increase in the number of re-id publications over the recent years. Finally, we discuss several challenging directions for future studies.

1. Introduction

Many countries consider video surveillance either as a primary tool to enforce security and prosecute criminals or simply as a crime deterrent tool. Following an incident, law enforcement authorities can review the available video footage, and identify a set of interest subjects, by matching the captured images/video to the enrolled IDs [9].

Given an input query, the person re-id systems compare and match the input data with the existing identities in the database (*gallery* set), probably captured from non-overlapping cameras and at different time intervals [3]. The goal is to retrieve an ordered list of the known identities with the most similarities to the query person. To this end, as outlined in Fig. 1(a), three modules (detection, tracking, and retrieval) work together, each one requiring a supervised learning phase on data that represent system settings. In the computer vision community, the tasks of person detection and tracking are considered independent fields that –at the end– help to obtain the gallery set. Therefore, aligned with the previous researches, in this paper we regard the person re-id exclusively as a retrieval problem that includes four main tasks: a) data collection; b) annotation; c) model training; and d) inference (see Fig. 1(b)).

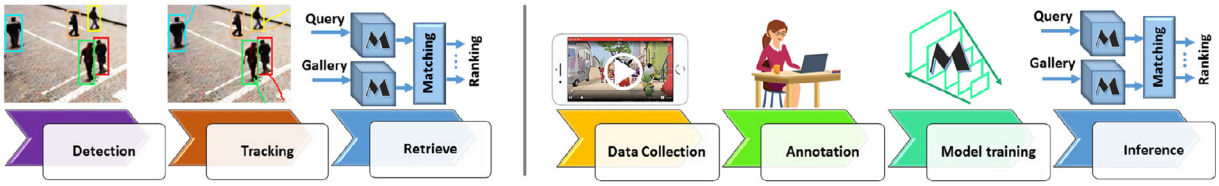
Full-body person re-id methods are either based on gait (dynamic) or appearance features. While gait is a *unique* behavioral biometric trait that is hard to counterfeit, it is highly dependent on the body-joints motion and can be affected by the slope of the surfaces, subjects' shoes and illness [25]. On the other side, appearance-based approaches rely on visual features such as edges, shape, color, texture, and expressiveness of the data. Therefore, being intrinsically different, the gait-based and visual-based approaches can be considered as disjoint tasks, both in terms of the existing databases and identification techniques. In this paper, for consistency purposes, we focus exclusively on the visual-based re-id approaches and refer the readers interested in gait-based re-id to [6,25].

Person re-id has attracted considerable interest in the last decade, with more than 53 papers published only in the CVPR 2019 and ICCV 2019 conferences. Over the past decade, many review articles have been published to organize the methods available in the research literature, each one study the problem from different and often contradictory perspectives. As relevant examples, Leng et al. [16] and Ye et al. [39] discuss the open-world setting versus close-world re-id and analyze the discrepancies, while [2,37] survey the methods from the deep learning point of view and emphasis the effectiveness of deep neural network structures upon re-id models performance. Wang et al. [36] addresses the challenge of heterogeneous re-id, in which the query and gallery sets allocate to different domains, and [23] studies the importance of efficiency and computational complexity in deep re-id architectures. Totally,

☆ Handle by Associate Editor Michele Nappi.

* Corresponding author.
E-mail address: Ehsan.yaghoubi@ubi.pt (E. Yaghoubi).

https://doi.org/10.1016/j.patrec.2020.12.017
0167-8655/



a) Person re-id system in operation (known as end-to-end) b) Steps for constructing a retrieval module from scratch

Fig. 1. An end-to-end re-id model detects and tracks the individuals in a video, and then retrieves the query person, while a typical re-id model focuses on the retrieval task.

we identified more than 20 body-based person re-id surveys, 12 were published as journal papers, 3 as books, and the remaining are available on ArXiv. From these resources, 9 papers have been published since 2019. For the complete list of surveys and reading more information about each article, we refer the readers to the Appendix.

1.1. Contributions

a) As our first and foremost motivation, we propose a multi-dimensional taxonomy that distinguishes between the person re-id models, based on their main approach, type of learning, identification settings, strategy of learning, data modality, type of queries and context (Section 2).
b) We briefly discuss the privacy and security concerns in surveillance, with a focus on Privacy-Enhancing Technologies (PETs), to encourage the research community to introduce privacy-by-design and default systems (Section 3).
c) We identify several emerging deviations caused by an evidently growing number of publications over the last few years and discuss the open issues and point out for future directions in this topic (Section 4).

However, the detailed analysis of the existing methods is out of the scope of our discussion, and this short survey of surveys should be regarded as a complement to the existing primary surveys.

2. Person re-identification taxonomy

Generally, re-id models have several independent features that help to categorize the methods from different perspectives, as shown in Fig. 2. Here, we not only provide a multi-dimensional taxonomy as an overall insight into the existing research, but we explore novel ideas from various points of view as well. As an example, the challenges in a deep learning model based on a text-query with open-world setting are totally different from the challenges of a model designed for a close-world setting with an RGB video-query. Therefore, in the following subsections, after discussing how data-acquisition and data-domain affect the re-id methods, we review the existing strategies for designing a re-id model, followed by a short description of the most popular approaches for implementation of the strategies. Finally, we briefly explain the categorization in system settings, context, data-modality, and learning-type.

2.1. Query-type

Before developing any re-id technique, two main properties of the data should be analyzed with particular attention:

2.1.1. Data-domain

In image-based datasets, the model is trained on a few samples per individual, while in video-based benchmarks, for each person, several *sequence of images* (i.e, video segments) are available. The existing video-based datasets consist of either RGB or infrared data [40], and both the query and gallery data are from the same domain (i.e., ***infrared-infrared, RGB-RGB***); whereas the image-based re-id datasets are classified into ***RGB-Depth, RGB-infrared, RGB-sketch, RGB-text***, and ***RGB-RGB***. RGB-RGB image-based datasets are classified into short-term and long-term re-id –in which identical persons may appear with ***different clothes***. When retrieving a person from a gallery, the operator may input a query that comes from different domains, which results in large distances between the features extracted from gallery and query data. When dealing with different data modalities, developing methods for learning the gap between domains is critical, since typical similarity features (e.g., texture and color) may be misleading.

2.1.2. Data-content

Data acquisition protocols and conditions (which could be performed either by handheld devices or stationary cameras) strongly determine the properties of the resulting data and affect the kind of re-id techniques suitable for the problem. For instance, as shown in Fig. 3, some data variability factors such as pose, motion, and occlusions heavily depend on the camera view angle and constraint the model's performance.

2.2. Strategies

Upon our analysis to the problem and to the existing surveys, we suggest that the existing re-id strategies can be broadly grouped according to five perspectives: scalability, pre-processing and augmentation, model architecture design, post-processing strategies, and robustness to noise.

2.2.1. Scalability

Speed, accuracy, and on-board processing are critical factors of a real-world person re-id system. The process of retrieving from large-size gallery sets is a time-demanding task, as a solution of which, designing ***efficient models*** and using ***hashing*** techniques have been effective. The unnecessary parts and parameters of the network are removed using pruning or distillation techniques [30] to increase the efficiency and build light-weighted models. Subsequently, the captured data can be processed on-board instead of transferring it to the operation center. Hashing [34] is the transformation of the features to a compressed form, which not only accelerates the searching process (matching) but occupies less area for storage as well. To tackle the problem of scalability in training phase and learn from huge volume of unlabeled data, a common solution is to apply ***transfer learning*** that is sometimes referred to as domain adaptation, in which we use an annotated source domain to learn the discriminative representation of the unlabeled target domain.

2.2.2. Pre-processing and augmentation

Apart from the basic pre-processing techniques (such as channel-wise color alteration or random erasing) that increase the volume of the labeled data, most of the methods in this category use Generative Adversarial Networks (GANs) to synthesize new

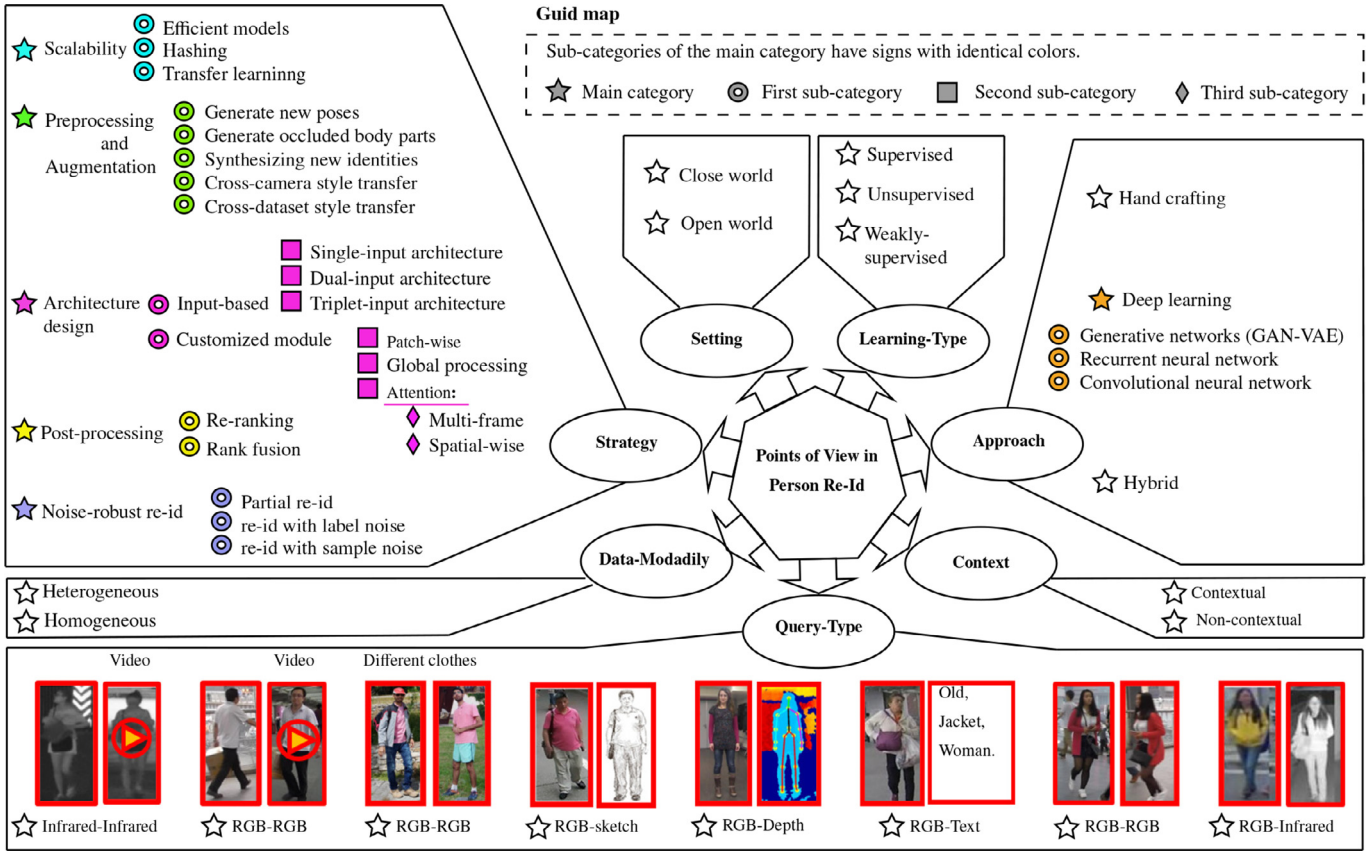


Fig. 2. Multi-dimensional taxonomy (*Points-of-view*) of the person re-identification problem.

Fig. 3. Examples of how varying capturing angles affect the salient points in the data and demand specific re-id solutions to obtain acceptable performance.

data or edit the existing ones. ***Generate new poses*** for the existing identities is a technique that allows the network to learn a comprehensive presentation of individuals, while ***generating occluded body-parts*** provides the model with new sets of features. Moreover, ***synthesizing new identities*** can be seen as a data augmentation technique that contributes to the re-id models' performance if the synthetic data follows a similar distribution to the original dataset.

The data undergoes substantial changes in color-style if we collect them from multiple cameras. However, a ***cross-camera style transfer*** can cross-transforms the color and illumination between cameras, which can strongly improve the model performance. Performing style transfer over multiple datasets (***cross-dataset style transfer***) is also used to increase the volume of the training data in the desired domain (e.g., transferring the style of night images to RGB images).

2.2.3. Architecture design

The quality of the extracted features from the query and gallery sets is a factor that significantly determines the system's performance. Generally, there are two overlapped perspectives to design a novel architecture for extracting discriminative representation from the data:

1) Design ***stream-based*** models, which could be investigated from two points of view: (a) in the first perspective, the main objective is to learn suitable metrics learning (using the loss function) to reduce the intra-class variations and increase the inter-class variations [15]. Different from typical re-id models that use ***single-stream architectures***, some novel models propose to use ***dual-stream architectures*** to focus on the inputs' similarity degree. Moreover, ***triplet/quadruplet-stream architectures*** use the images of the other identities as negative inputs and the images of the target person as positive and anchor inputs [14]. It worth mentioning that usually the weights and parameters are sheared between streams of the model, leading to a popular architecture called Siamese networks [29]. (b) the second perspective to design stream-based models is to extract various features from one identity using multiple streams and fuse them together (e.g., fusing extracted information from motion,

Fig. 4. Some of patching strategies used to obtain fine-grained local representations of the input data.

semantic attributes, handcrafting techniques, and CNN-based methods).

2) Design ***customized modules*** to perform specific processes for extracting robust discriminative features from data. When discussing the customized-design, there are many possibilities; therefore, we sub-categorize them into three groups: (a) ***Patch-wise techniques***. Patch-based analysis helps to extract minutiae information (known as fine-grained features) from the data, which helps to discriminate between inter-class samples that are visually similar to each other. Not only can the patch-wise techniques use various ways of patching (illustrated in Fig. 4), but they use different approaches to analyze each patch as well. For example, when using a simple Long Short-Term Memory (LSTM) architecture, the comprehensive feature representation is obtained by processing all the patches one after another, while in a multi-input architecture, one can perform a cross-analysis –e.g., to extract shareable features from head-patches of two images. (b) ***Global-based processing techniques*** focus on the topology of the cameras and network consistency [34]. Three widely-used datasets (i.e., Market1501, DukeMTMC, and GRID) have provided the locations (aerial map), where each camera covers, to allow studying the effects of cameras' topology on the model efficiency. As a vivid example, suppose two cameras cover the entrance and exit sides of a narrow street; thus, a person that is firstly captured in frontal-view probably appears in rear-view on the next camera. (c) ***Attention-based techniques***. By capturing images from different angles, some parts of the input-data undergo substantial changes in appearance, texture, shape, occlusion, and illumination. Fundamentally, this is a misalignment problem, in which the model aims to find the target person by matching the corresponding regions of the body (e.g., head with head) in query and gallery data. The existing solutions are typically divided into: (i) ***special-wise*** attention; and (ii) ***multi-frame*** attention. Generally, special-wise techniques search for salient pixels/regions on the image, which could be accomplished by performing a channel-wise operation, learning hard-masks, developing modules for regional selection or by designing multi-input networks. In the multi-frame attention architecture, the aim is to provide *one* feature representation from a *sequence of images*.

2.2.4. Post-processing

The output of a re-id model is an ordered list of gallery identities, according to the similarity between the gallery and query data. This list is called ranking-list, and any further processes for re-ordering the results are known as ***re-ranking***. Many intuitive scenarios could help refine this ranking list. For example, in case of being ranked particularly high for one query, a gallery image should be ranked low for any other queries. Also, if the query person has dark-skin, individuals with light-skins should not be ranked high. Another frequent post-processing approach is the *rank fusion* (fusion of ranking-lists) of multiple re-id methods, which is particularly suitable when accuracy is much more important than speed and computational cost.

2.2.5. Robustness to noise

Whether we use automatic human detection and tracking or perform it manually, errors, misalignment, and inconsistency in bounding-box detection are inevitable. Furthermore, the annotation process is a human-biased step that is mostly accompanied by some percentage of errors that may affect the quality of the learning process. There are three general approaches to tackle these challenges [39]. ***Partial re-id*** techniques construct models capable of extracting shareable features from unoccluded body parts, while outlaid bounding boxes and inaccurate tracking are studied under the ***sample-noise reduction. Label-noise*** topic addresses the annotation errors by limiting the model not to be overfilled on the labels.

2.3. Approaches

The discussed strategies (in Section 2.2) could be taken in to account by three approaches: ***deep learning, hand-crafting***, and the combination of both (***hybrid***).

In the last decade, re-id systems were usually implemented based on knowledge-based feature extractors, which could be classified into four main groups: camera geometry/calibration, color calibration, descriptor learning, and distance metric learning. As most of the traditional techniques were built upon appearance-based similarities, designing discriminative visual descriptors and learning distance metrics upon person clothes were more popular than other methods [32].

Many studies focused on deep structures or a combination of deep neural networks and traditional methods after the advent of deep learning approaches. In the context of deep learning, ***Convolutional Neural Networks*** (CNNs) analyze the input data at a single instance, while in ***Recurrent Neural Networks*** (RNNs) the data is treated as a sequence of inputs; then, taking advantage of an internal state (memory), the critical information of each sequence is accumulated to construct the final feature representative of the input. Finally, ***generative networks*** are classified into Variational Auto-Encoder (VAE) and GAN, each aiming to find the distribution of the original dataset to generate new data. In re-id, GAN-based approaches have shown promising results with both augmenting the dataset, and editing the samples (e.g., style transferring, completing the occluded body-parts, etc.)

2.4. Identification settings

Re-id model are either classified into the ***open-world*** or ***closed-world*** settings. The closed-world assumption deals with matching one-to-many samples, so that the query image is surely corresponding to one of the gallery individuals. On the other hand, there are different interpretations for the open-world setting: (1) it might regard a multi-camera problem in which the gallery evolves over time, and the ever-changing query may not be presented in the gallery. Moreover, the system could re-identify multi-subjects at once [31]; (2) it might regard a group-based verification task aiming to determine whether the query appears in the gallery or

not, without the necessity of retrieving matched person(s) [4]; and (3) any real-world application that excludes the close-world setting could be considered as an open-world problem. For example, in Ye et al. [39], researches that deal with heterogeneous data, raw images/videos, limited labels, and noisy annotations have been considered as open-world studies [35].

2.5. Context

Context is another point of view towards re-id problems so that if the system relies on the external contextual information (e.g., camera/geometric information) rather than using the data itself, it is considered as a contextual system [3]. However, after the advent of deep learning technologies, only a small proportion of works consider person re-id from the contextual perspective [34]. Meanwhile, contextual based re-id datasets should provide extra information such as full-frame data, cameras' locations and capturing angles e.g., using an aerial map.

2.6. Data-modality

Given the various data modalities for the query and gallery sets, the re-id task can be regarded either as a heterogeneous re-id (He-Reid) or homogeneous (Ho-Reid) problem. In a Ho-Reid perspective, the query and gallery data have similar modalities, while in the He-Reid the query is from another domain (for example, if the gallery consists of RGB-images, the query could be a verbal description of the target person). Therefore, in He-Reid, discrepancies between the query-domain and the gallery-domain are huge so that the methods developed for Ho-Reid cannot be directly applied to these problems. Dealing with two different data modalities, He-Reid techniques aim to bridge the gap between domains and decrease the inter-modality discrepancy, for which there are several methods [36]: (1) learning a metric to decrease the gap between features of each domain; (2) learning shared features; and 3) unifying modalities before feature extraction by transferring both domains to a latent domain. So far, owing to Generative Adversarial Networks (GAN), unifying the modalities has shown better results that are discussed at the end of this section.

2.7. Learning-type

Supervised, semi supervised, weakly supervised, and unsupervised learning [20] are the annotation-based learning types. Due to leveraging the manually annotated data, ***supervised*** methods achieve superior accuracy than other methods. However, some works develop ***weakly-supervised*** or ***unsupervised methods*** to not only ease the process of data annotation but also train the model on an excessive amount of unlabeled data. The main categories in unsupervised learning are *domain adaption, dictionary learning, feature representation extraction, distance measurement*, and *clustering* [34], from which *Unsupervised Domain Adaptation* (UDA) has attracted the most attention. In UDA, taking advantage of a labeled dataset (source domain), the model learns the discriminative representation of the unlabeled data (target domain). Therefore, the distance between the data distribution of domains is minimized, so that target-domain data can be treated as the source-domain data for training purposes. Different from the time-consuming annotation process for supervised methods (all people in the video are annotated one-by-one), weakly-supervised annotation is a video-level process, in which each video needs one label, indicating the IDs appeared in that video.

2.8. State-of-the-art performance comparison

Table 1 shows the performance (rank-1 and mean Average Precision (mAP)) of the state-of-the-art techniques, most published in 2019 and 2020. In these works, the gallery set is always composed of RGB images/videos, except in Zhang et al. [40] (with 14.3% accuracy for rank-1 retrieval), where both the gallery and query sets contain infrared images captured at night.

Table 1
Performance of the state-of-the-art re-id methods.

Field of study	Dataset	Method	R 1	mAP
RGB-Thermal	RegDB	[39]	70.0	66.4
RGB-infrared	SYSU-MM01	[17]	49.9	50.7
RGB-Sketch	Sketch Re-ID	[10]	49.0	–
RGB-Text	CUHK-PEDES	[1]	56.7	–
Infrared-infrared	KnightReid	[40]	14.3	10.2
RGB-D	KinectReID	[28]	99.4	–
	RGBD-ID	[28]	76.7	–
Unsupervised	Market-1501	[8]	86.2	68.7
	DukeMTMC*	[8]	76.0	60.3
RGB image-based	Market-1501	[5]	95.7	89.0
	CUHK03	[39]	63.6	62.0
	MSMT17	[39]	68.3	49.3
	DukeMTMC*	[5]	91.1	81.4
RGB video-based	3DPeS	[42]	78.9	–
	PRID2011	[18]	95.5	–
	iLDS-VID	[19]	88.9	93.0
	MARS	[21]	90.0	82.8
	DukeMTMC-VideoReID*	[19]	96.2	95.4
	LS-VID	[18]	63.1	44.3
	PRW	[38]	73.6	33.4
Long-term	Motion-ReID*	[41]	65.7	–
	Celeb-reID	[12]	51.2	9.8

*Not publicly available.

Wang et al. [36] reported that the performance of all the He-Reid works is lower than 40%, whereas the latest papers have claimed 56.7%, 49.0%, 49.%9, and 70.0% rank-1 accuracy for RGB-text, RGB-sketch, RGB-infrared, and RGB-thermal, respectively, pointing for a fast improvement in performance in this field.

Table 1 enables to conclude that He-Reid and long-term re-id are the least matured fields of study, respectively with 70% [39] and 65.7% [41] rank-1 accuracy, while [8] is an unsupervised person re-id work that is close to the hopeful boundary, with 86.2 % and 76% rank-1 accuracy on the Market-1501 and DukeMTMC datasets, respectively.

On the other hand, even though studies based on RGB images and RGB videos have achieved higher results, their performance is highly dependent on the dataset, such that rank-1 accuracy in RGB video-based studies is in a rage from 63.1 % [18] to 96.2% [19] for the LS-VID and DukeMTMC-VideoReID datasets, respectively; similarly, in RGB image-based researches, Chen et al. [5] has achieved 95.7% rank-1 accuracy on the Market-1501 dataset, while [39] reports this number around 63.6% for their experiments on the CUHK03 dataset.

3. Privacy concerns

IAPP, the International Association of Privacy Professionals, defines that *privacy* is the right to be free from interference or intrusion and to remain anonymous, and *information privacy* regards the control over our own personal information. Among the possible ways of privacy violation [27] (i.e., watching, listening, locating/tracking, detecting/sensing, personal data monitoring, and data analytics), we pay attention to the visual monitoring that has recently engaged the research community, due to the sensitiveness of monitoring people or collecting their personal visual data (from the Internet) without their consent. In this scope, several well-known benchmarks (e.g., *Brainwash, DuckMTMC*, and *MS-Celeb-1M*) were permanently suspended by their authors [11], in most cases due to the absence of explicit authorization from the subjects in

the dataset to have their data collected and disseminated for research purposes.

Overall, there are two solutions to reduce the privacy concern in person re-id models: ***privacy-by-design principles*** and ***Privacy-Enhancing Technologies*** (PETs).

Privacy-by-design principles are some standards to protect data through technology design, published by the law enforcement agencies[1,2] and enforce companies to respect the privacy of their customers. In these standards, information tracking is defined as a principle that allows people to manage and track whom they have access to their private information (and to what extend). In contrast, the data minimization principle states that enterprises should only process the minimum necessary data. For example, a visual surveillance panel that processes the crowd for displaying related advertisements may need to recognize the human semantic attributes (e.g., gender, clothing styles, etc.), but should avoid designing a system that detects faces, analyzes the skin color, and people's race.

PETs are methods of protecting data, including anonymization, perturbation, and encryption [7]. In anonymization, the sensitive information is removed to perform a complete de-identification, generally accomplished by masking, while in perturbation, the sensitive attributes of the data are replaced with noisy or otherwise altered data. On the other hand, security techniques *reversibly* disguise the identifying information. Examples of PETs in person re-id could be disguising pedestrian's faces in the gallery set using generative networks to reduces the risk of privacy intrusion; however, it indicates the need for methods that are able to perform the re-id task on anonymized data and possibly reconstruct the true faces if asked by the authorities [26]. As a method for developing fast re-id, *hashing* could be used to design a re-id model that works with encrypted data and reduces the risk of hacking.

4. Discussion and future directions

4.1. Biases and problems

The number of methods in person re-id has considerably increased in recent years, leading to some biases and problems such as unfair comparisons, low originality in techniques, and insufficient attention to some of the important perspectives in the problem.

4.1.1. Unfair comparisons

Based on the re-implementation of several state-of-the-art re-id methods, a recent baseline [22] explicitly concluded that the improvements reported in some works were mainly due to training tricks rather than to any conceptual advancement of the method itself, which has led to an exaggeration of the success of such techniques. Therefore, to show the effectiveness of the model, we suggest to perform an ablation study on the proposed method, such that the basic model is first evaluated, and each proposed component is added one by one over the baseline to show the effectiveness of the idea. Further, to show the superiority of the method over the existing state of the arts, authors should remain the architecture and parameters constant as much as possible, so that we are certain that the improvement is caused by the idea [24].

4.1.2. Low originality

Although using the power of other fields in person re-id is valuable and improves the performance of state of the art, in recent years, excessive attention to these kinds of contributions has decreased the number of original works with significant contributions. In the literature, we repeatedly face with re-implementation of other fields' ideas as original re-id methods, creating competition for a mere copy of outside ideas into re-id problems. For example, as confirmed by Musgrave et al. [24], after the success of LSTM, GAN, Siamese network, backbone networks (ResNet, Inception, GoogleNet), various loss functions, etc., many authors repeated the same ideas on the re-id datasets.

4.1.3. Insufficient attention to some perspectives

A long-term re-id model capable of retrieving multi-modality queries is much more realistic and useful than a close-world, single-modality retrieval system. Nevertheless, why does exist more researches in the second scenario?. Understanding the nature of the deep neural network is the answer to this question. It is known that deep neural networks are efficient in feature extraction, and they have shown promising results specifically in problems dealing with appearance-based features. Thereby, re-id scenarios under close-world setting and homogeneous RGB data-modality have shown considerable performance improvement. On the other hand, there is little attention to some challenges such as open-world setting, long-time re-id, heterogeneous modality, and non-contextual tasks.

4.2. Open issues

In this section, we discuss the major open issues in the re-id problem and point out for some possible further directions.

Person re-id performance has several important covariates, such as variations in background, illumination, occlusion, body-pose, and other view-dependent variables [16,33]. In particular, we emphasize the role of data annotation: when training deep neural networks, the more the annotated data are available, the better the performance would be. However, data preparation for re-id is an expensive, tedious, and time-consuming process, opening the space for developing novel semi-supervised, weakly supervised or even unsupervised solutions for training the models [39].

Affected by similar covariates, other pattern recognition tasks (e.g., iris recognition, cross-domain clothing analysis, multi-object tracking) have significantly helped the person re-id in several directions such as unsupervised learning, extraction of discriminative feature sets, and application of robust metric learning techniques. Nevertheless, some challenges are related explicitly to the re-id task: for example, by increasing the volume of the re-id datasets, the matching process (for retrieving the query person from a large-scale gallery set) takes substantially more time, indicating the need for fast re-id methods [39].

Furthermore, for a real-world re-id system, it is necessary to search the query person independent of its data-type. However, due to the lack of large datasets consisted of multi-modal data, current heterogeneous works are limited to single cross-modality searches, and the gallery set often consists of RGB images. Unifying modalities of the query set and gallery set is another open issue in heterogeneous re-id that could be fulfilled by mapping the modality of both sets either to each other interchangeably or to a latent space [36].

Apart from most of researches in the literature that are based on appearance, long-term re-id solves the issue of retrieving the same person with different appearance and clothing style [2]. Therefore, studies in this area should consider challenges such as: (1) going beyond appearance-based features and extract discriminative features from hard-biometrics (face and gait) and more robust soft-biometrics (height, body volume, body contours). Meanwhile, recent facial recognition techniques that typically are trained on high-resolution data (with controlled pose-variation)

[1] https://gdpr-info.eu/.
[2] https://www.priv.gc.ca/en/report-a-concern/.

may not increase the overall performance when dealing with low-quality faces in the wild; (2) long-term re-id in real applications is often tied to open-world setting challenges such as scalability (how to deal with large databases) and generalization (adding new cameras to the existing system) [16]. It worth mentioning that person search is a slightly different research area that aims to locate the prob person within a whole frame containing one/several persons [13].

Currently, a plethora of human detection and tracking techniques are available for different platforms. By generalizing them for the handheld devices –thanks to high-speed internet connections–, mobile person re-id can quickly become a trivial task, which raises many privacy and security concerns. Thus, both secure storage of the gallery set and proposing re-id methods that conform privacy concerns *by design and default* are of the utmost challenges.

5. Conclusion

Person re-id aims to retrieve an ordered list of the identities from a database, with respect to query images taken from one or multiple non-overlapping cameras. In result of the extensive research carried out over the last years for solving the primary pattern recognition challenges (e.g., pose variations, partial occlusions and dynamic data acquisition conditions), re-id systems have successfully passed the human accuracy-level in easy scenarios (i.e., when the model is trained based on supervised learning and close-world setting in RGB heterogeneous modality). In this paper, we proposed a multi-view taxonomy that considers the different categorizations available in the re-id literature to ease the discovery of realistic and feasible scenarios for future directions. Furthermore, we discussed the importance of the concept of privacy in this field and briefly reviewed several strategies to improve systems' security and privacy by default. Finally, after discussing some of the issues caused by an evidently growing number of publications in recent years, we pointed out for some of the open issues in this extremely challenging problem.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the FCT/MEC through National Funds and Co-Funded by the FEDER-PT2020 Partnership Agreement under Project UIDB/50008/2020, Project POCI-01-0247-FEDER-033395 and in part by operation Centro-01-0145-FEDER-000019 - C4 - Centro de Competências em Cloud Computing, co-funded by the European Regional Development Fund (ERDF) through the Programa Operacional Regional do Centro (Centro 2020), in the scope of the Sistema de Apoio à Investigação Científica e Tecnológica - Programas Integrados de IC&DT.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patrec.2020.12.017.

References

[1] S. Aggarwal, V.B. Radhakrishnan, A. Chakraborty, Text-based person search via attribute-aided matching, in: The IEEE Winter Conference on Applications of Computer Vision, 2020, pp. 2617–2625.
[2] M.O. Almasawa, L.A. Elrefaei, K. Moria, A survey on deep learning-based person re-identification systems, IEEE Access 7 (2019) 175228–175247.
[3] A. Bedagkar-Gala, S.K. Shah, A survey of approaches and trends in person re-identification, Image Vis. Comput. 32 (4) (2014) 270–286.
[4] S. Chan-Lang, Closed and Open World Multi-shot Person Re-identification, Paris 6, 2017 Ph.D. thesis.
[5] H. Chen, B. Lagadec, F. Bremond, Learning discriminative and generalizable representations by spatial-channel partition for person re-identification, in: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2472–2481.
[6] P. Connor, A. Ross, Biometric recognition by gait: a survey of modalities and features, Comput. Vis. Image Underst. 167 (2018) 1–27.
[7] J. Curzon, A. Almehmadi, K. El-Khatib, A survey of privacy enhancing technologies for smart cities, Pervasive Mob. Comput. 55 (2019) 76–95, doi:10.1016/j.pmcj.2019.03.001.
[8] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, T.S. Huang, Self-similarity grouping: a simple unsupervised cross domain adaptation approach for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6112–6121.
[9] M. Gou, Z. Wu, A. Rates-Borras, O. Camps, R.J. Radke, et al., A systematic evaluation and benchmark for person re-identification: features, metrics, and datasets, IEEE Trans. Pattern Anal. Mach. Intell. 41 (3) (2018) 523–536.
[10] S. Gui, Y. Zhu, X. Qin, X. Ling, Learning multi-level domain invariant features for sketch re-identification, Neurocomputing 403 (2020) 294–303, doi:10.1016/j.neucom.2020.04.060.
[11] A. Harvey, J. LaPlace, Megapixels: origins, ethics, and privacy implications of publicly available face recognition image datasets, (2019). https://megapixels.cc/ (accessed January 7, 2021).
[12] Y. Huang, J. Xu, Q. Wu, Y. Zhong, P. Zhang, Z. Zhang, Beyond scalar neuron: adopting vector-neuron capsules for long-term person re-identification, IEEE Trans. Circuits Syst. Video Technol. 30 (10) (2019) 3459–3471.
[13] K. Islam, Person search: new paradigm of person re-identification: a survey and outlook of recent works, Image Vis. Comput. 101 (2020) 103970, doi:10.1016/j.imavis.2020.103970.
[14] A. Khatun, S. Denman, S. Sridharan, C. Fookes, A deep four-stream siamese convolutional neural network with joint verification and identification loss for person re-detection, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2018, pp. 1292–1301.
[15] B. Lavi, I. Ullah, M. Fatan, A. Rocha, Survey on reliable deep learning-based person re-identification models: Are we there yet?, arXiv:2005.0035 (2020).
[16] Q. Leng, M. Ye, Q. Tian, A survey of open-world person re-identification, IEEE Trans. Circuits Syst. Video Technol. 30 (4) (2020) 1092–1108.
[17] D. Li, X. Wei, X. Hong, Y. Gong, Infrared-visible cross-modal person re-identification with an x modality, in: Proc. AAAI Conf. Artif. Intell., 2020, pp. 4610–4617.
[18] J. Li, J. Wang, Q. Tian, W. Gao, S. Zhang, Global-local temporal representations for video person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3958–3967.
[19] M. Li, H. Xu, J. Wang, W. Li, Y. Sun, Temporal aggregation with clip-level attention for video-based person re-identification, in: The IEEE Winter Conference on Applications of Computer Vision, 2020, pp. 3376–3384.
[20] Y. Lin, Deep learning approaches to person re-identification, University of Technology Sydney, 2019 Ph.D. thesis.
[21] C.-T. Liu, C.-W. Wu, Y.-C. F. Wang, S.-Y. Chien, Spatially and temporally efficient non-local attention network for video-based person re-identification, arXiv:1908.01683 (2019).
[22] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, J. Gu, A strong baseline and batch normalization neck for deep person re-identification, IEEE Trans. Multimed. 22 (10) (2019) 2597–2609, doi:10.1109/TMM.2019.2958756.
[23] H. Masson, A. Bhuiyan, L.T. Nguyen-Meidine, M. Javan, P. Siva, I.B. Ayed, E. Granger, A survey of pruning methods for efficient person re-identification across domains, arXiv:1907.02547 (2019).
[24] K. Musgrave, S. Belongie, S.-N. Lim, A metric learning reality check, arXiv:2003.08505 (2020).
[25] A. Nambiar, A. Bernardino, J.C. Nascimento, Gait-based person re-identification: a survey, ACM Comput. Surv. 52 (2) (2019), doi:10.1145/3243043.
[26] H. Proença, The uu-net: Reversible face de-identification for visual surveillance video footage, arXiv:2007.04316 (2020).
[27] C. D., Raab, Privacy, security, surveillance and regulation, (2017). http://www.inf.ed.ac.uk/teaching/courses/pi/2017_2018/slides/RaabProfIssuesInformaticsCourse2017FINALppt.pdf (accessed January 7, 2021).
[28] L. Ren, J. Lu, J. Feng, J. Zhou, Uniform and variational deep learning for RGB-D object recognition and person re-identification, IEEE Trans. Image Process. 28 (10) (2019) 4970–4983.
[29] S.K. Roy, M. Harandi, R. Nock, R. Hartley, Siamese networks: the tale of two manifolds, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3046–3055.
[30] I. Ruiz, B. Raducanu, R. Mehta, J. Amores, Optimizing speed/accuracy trade-off for person re-identification via knowledge distillation, Eng. Appl. Artif. Intell. 87 (2020) 103309, doi:10.1016/j.engappai.2019.103309.
[31] M.A. Saghafi, A. Hussain, H.B. Zaman, M.H.M. Saad, Review of person re-identification techniques, IET Comput. Vis. 8 (6) (2014) 455–474.
[32] R. Satta, Appearance descriptors for person re-identification: a comprehensive review, arXiv:1307.5748 (2013).

[33] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, S.Z. Li, Embedding deep metric for person re-identification: astudy against large variations, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham, 2016, pp. 732–748.
[34] H. Wang, H. Du, Y. Zhao, J. Yan, A comprehensive overview of person re-identification approaches, IEEE Access 8 (2020) 45556–45583.
[35] H. Wang, X. Zhu, T. Xiang, S. Gong, Towards unsupervised open-set person re-identification, in: 2016 IEEE International Conference on Image Processing (ICIP), IEEE, 2016, pp. 769–773.
[36] Z. Wang, Z. Wang, Y. Wu, J. Wang, S. Satoh, Beyond intra-modality discrepancy: a comprehensive survey of heterogeneous person re-identification, arXiv:1905.10048 (2019).
[37] D. Wu, S.-J. Zheng, X.-P. Zhang, C.-A. Yuan, F. Cheng, Y. Zhao, Y.-J. Lin, Z.-Q. Zhao, Y.-L. Jiang, D.-S. Huang, Deep learning-based methods for person re-identification: a comprehensive review, Neurocomputing 337 (2019) 354–371, doi:10.1016/j.neucom.2019.01.079.
[38] Y. Yan, Q. Zhang, B. Ni, W. Zhang, M. Xu, X. Yang, Learning context graph for person search, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2158–2167.
[39] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, S.C. Hoi, Deep learning for person re-identification: a survey and outlook, arXiv:2001.04193 (2020).
[40] J. Zhang, Y. Yuan, Q. Wang, Night person re-identification and a benchmark, IEEE Access 7 (2019) 95496–95504.
[41] P. Zhang, Q. Wu, J. Xu, J. Zhang, Long-term person re-identification using true motion from videos, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2018, pp. 494–502.
[42] S. Zhou, J. Wang, D. Meng, Y. Liang, Y. Gong, N. Zheng, Discriminative feature learning with foreground attention for person re-identification, IEEE Trans. Image Process. 28 (9) (2019) 4671–4684.