Contents lists available at ScienceDirect

Image and Vision Computing

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

Person search: New paradigm of person re-identification: A survey and outlook of recent works

Khawar Islam

FloppyDisk.AI, Karachi, Pakistan

ARTICLE INFO

Article history:
Received 18 June 2020
Accepted 24 June 2020
Available online 30 June 2020

Keywords:
Person re-identification
Person search
Literature survey
Metric learning
Loss functions

ABSTRACT

Person Search (PS) has become a major field because of its need in community and in the field of research among researchers. This task aims to find a probe person from whole scene which shows great significance in video surveillance field to track lost people, re-identification, and verification of person. In last few years, deep learning has played unremarkable role for the solution of re-identification problem. Deep learning shows incredible performance in person (re-ID) and search. Researchers experience more flexibility in proposing new methods and solve challenging issues such as low resolution, pose variation, background clutter, occlusion, viewpoints, and low illumination. Specially, convolutional neural network (CNN) achieves breakthrough performance and extracts useful patterns and characteristics. Development of new framework takes substantial efforts; hard work and computation cost are required to acquire excellent results. This survey paper includes brief discussion about feature representation learning and deep metric learning with novel loss functions. We thoroughly review datasets with performance analysis on existing datasets. Finally, we are reviewing current solutions for further consideration.

1. Introduction

As an essential and demanding problem in computer vision, person re-ID [1,2] and person search [3] have emerged an independent topic and fast-growing topic in computer vision that deals with person retrieval in videos and digital images. Deep learning has become a major technique for researchers; the victory of deep learning methods has conducted new wave into person re-identification, moving towards to a research highlight. Person (re-ID) has been broadly used in academic community and large-scale industry implementation, such as public safety, tracking of person in widespread public parks, universities and streets, behavior analysis and surveillance. From the perspective of video surveillance community, the energizing and most critical issue of person search is to correctly match a probe person that has been observed in cameras at different locations under heavy intensive changes in pose, viewpoints and lighting has more significance importance in research community.

There are mostly two major techniques of person retrieval. One method is to match an investigated person with the gallery of manually cropped persons which is little far from real-world applications (see Fig. 1). While person search is an advanced and exacting task which seeks to identify person from whole scene of images which is closer to real world implementation. Another technique is person search (PS) detects all persons and recognize the probe person in an image exclusive of proposal and bounding box that is divergent from person re-ID (see Fig. 2). Nevertheless, person search is near to physical world and quite demanding but as automated pedestrian's detection makes incorrect detection and unbalanced images. That's why: solving complex problems of person re-ID and person search are extremely hard job due to human posture, light, video camera position, low resolution, occlusion, viewpoints, pose estimation and different variations in images [5–10]. Biometric signs like face and walking style are impossible to detect and track a person in low resolution of cameras [11,12]. The re-identification and person search task only rely on visual appearance [13,17]. Although several detectors achieve well on lighting and occlusion free situation, some of them produce imperfectly on the task of detecting person. This is because of above aforementioned issues in person search. Moreover, the knowledge of person search is very little because previous accomplishment is focused on person (re-ID) problems. (See Fig. 3.)

This paper proposes an extensive and in-depth review on person search. Our survey objective is to make comprehensively six parts of person search methods, including feature learning, architectural design, deep metric learning, and loss function design. Apart from taxonomically discussing the previous person search techniques, we extensively explore previous datasets of person search. We comprehensively examine the performance of person search techniques and present ongoing promising directions for future work.

E-mail address: khawar512@gmail.com.

https://doi.org/10.1016/j.imavis.2020.103970

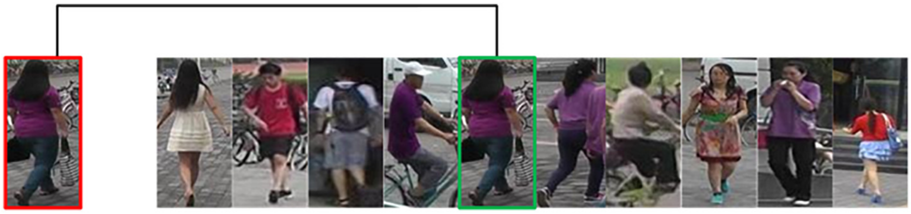

Fig. 1. Simplified view of person re-identification (re-ID): full scene images were taken from different source of camera, after cropped all persons from images then match an image of manually cropped person.

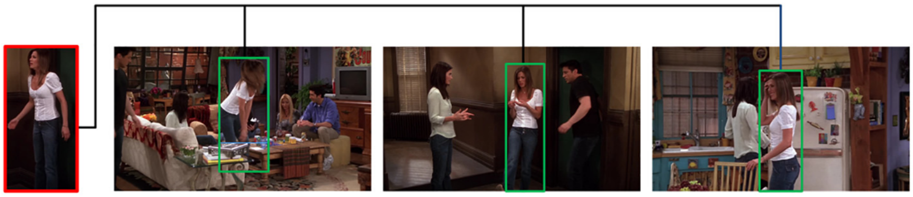

Fig. 2. Overview of person search: finding person from whole scene of images without proposal and bounding boxes.

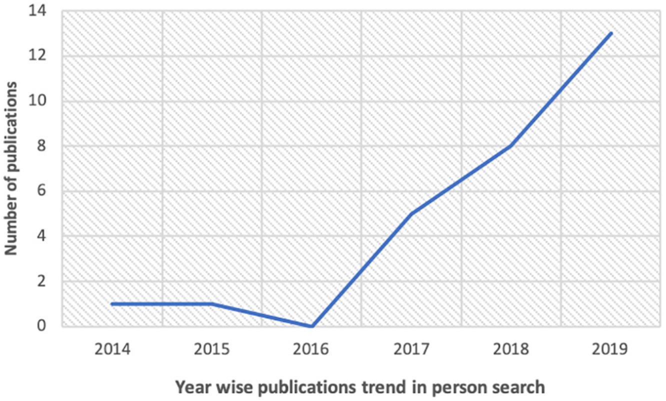

Fig. 3. Number of research publications on top conferences and journals over recent years. Year wise distribution shows the significant importance in last two years.

1.1. Background and scope

The history of person search is comparatively little. Recent work on person search is mainly about person retrieval using hand-crafted features and deep learning techniques in images [4,11]. Before the widespread of deep learning, commonness in shape and human body appearance-based features are utilized [11]. Upon the rapid progression of CNN, various deep learning-based person search approaches have emerged. However, mostly surveys and researches emphasizing hardly on person (re-ID). In order to take first step, Yuanlu et al. [11] put forward the first step to introduce the person search uniqueness, a Gaussian Generative Model (GMM) to catch commonness between persons, and provide a baseline result in order to investigate and explore more about person (re-ID). Later, Tong et al. [4] proposed and suggested an effective person search technique which cooperatively handles person detection and (re-ID) in same convolutional neural network. Completely different from the GMM and Fisher vector used in [11], Jinfu Yang et al. [31] use hand crafted features [14] for pedestrian detection and retrieval in images. SSD [72] and faster r-cnn [65] are two major detectors in pedestrian detection. Some pedestrian detection approaches [4,40,54–56,59] utilize Faster R-CNN, some researchers modify [65] and proposed more accurate pedestrian detector [62]. Furthermore, feature learning [20,22,31], multi-scale feature learning [3,30,56], architecture designs [4,31], loss functions [54–62], major contribution [21,22] and convolutional neural networks [48,49] are also utilized for person search. A brief timeline illustrated in Fig. 4.

We have precisely and thoroughly selected remarkable and distinguished papers published in outstanding journals and top venue conferences. This survey focuses on the rapid advancement in person search in last four to six years. In addition to this, some other correlated work is also comprised to make our work more understanding and helpful in this field. We restrict this survey paper only for person search techniques and little discussion about recent work in person (re-ID). Other than that, some work related to person (re-ID), such as re-identification in two separate steps will also be included in our discussion.

1.2. Comparison with previous surveys

Several remarkable person (re-ID) surveys have been published. The particular comprise numerous reviews on the person retrieval under certain conditions, such as heterogeneous person (re-ID) [73,78], open-world and closed-world settings [15,74], hand-crafted features and deep learning techniques for image and video-based [74], traditional approaches [76], and gait based person (re-ID) [7,77]. In addition to these subject specific person (re-ID) surveys and reviews, numerous numbers of general person (re-ID) surveys [1,10,15,16,75,79]. Among these, Srikrishnaet al. [1] extensively evaluate performance of single and multi-shot (re-ID) with metric learning, including feature extraction, ranking techniques, evaluate methods on new large-scale dataset and 16 publicly available datasets. Bahramet al. [75] gives comprehensive survey on person (re-ID) using deep neural networks: he mentioned some guidelines and limitations. Moreover, Mohammad et al. [79] briefly discussed many aspects of person (re-ID), covering intra/inter camera, spatial features, textual and appearance descriptors, and datasets. Similarly, Di et al. [80] thoroughly reviewed many deep models, including identification, verification, deep metric learning, video-based and data augmentation based deep models. Furthermore, he broadly reviewed previous person (re-ID) in several years, datasets and discussed future insights. From the viewpoint of person search (re-ID), these surveys do not discuss and focus on person search. We are introducing a systematic and complete review of deep learning algorithms that mutually handle person detection and re-ID. This survey is focused by critical and in-depth analysis of person search. We sum up existing person search algorithms based on six different angles: feature

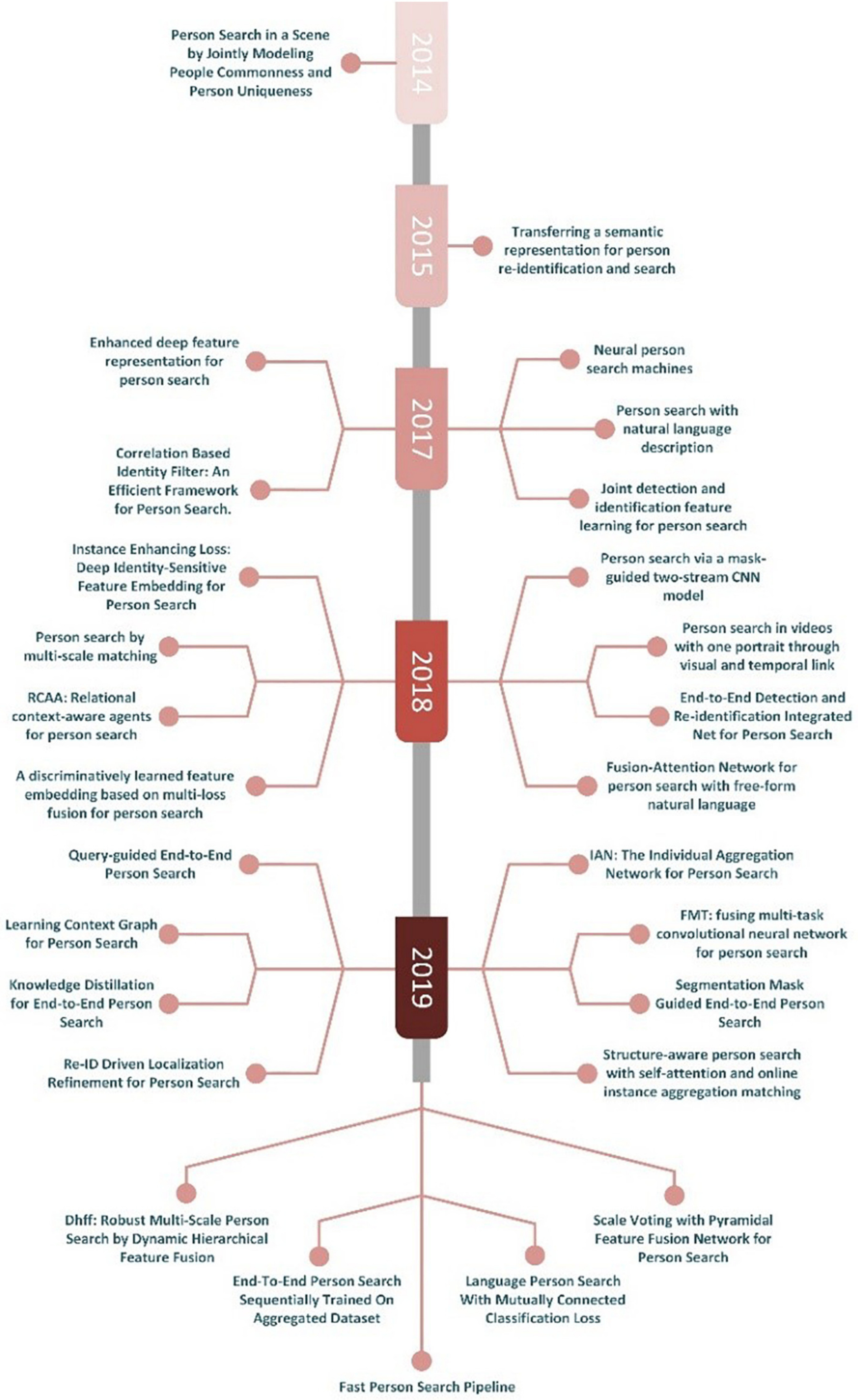
2014
Person Search in a Scene by Jointly Modeling People Commonness and Person Uniqueness
2015
Transferring a semantic representation for person re-identification and search
2017
Enhanced deep feature representation for person search
Correlation Based Identity Filter: An Efficient Framework for Person Search.
Neural person search machines
Person search with natural language description
Joint detection and identification feature learning for person search
2018
Instance Enhancing Loss: Deep Identity-Sensitive Feature Embedding for Person Search
Person search by multi-scale matching
RCAA: Relational context-aware agents for person search
A discriminatively learned feature embedding based on multi-loss fusion for person search
Person search via a mask-guided two-stream CNN model
Person search in videos with one portrait through visual and temporal link
End-to-End Detection and Re-identification Integrated Net for Person Search
Fusion-Attention Network for person search with free-form natural language
2019
Query-guided End-to-End Person Search
Learning Context Graph for Person Search
Knowledge Distillation for End-to-End Person Search
Re-ID Driven Localization Refinement for Person Search
IAN: The Individual Aggregation Network for Person Search
FMT: fusing multi-task convolutional neural network for person search
Segmentation Mask Guided End-to-End Person Search
Structure-aware person search with self-attention and online instance aggregation matching
Dhff: Robust Multi-Scale Person Search by Dynamic Hierarchical Feature Fusion
End-To-End Person Search Sequentially Trained On Aggregated Dataset
Language Person Search With Mutually Connected Classification Loss
Scale Voting with Pyramidal Feature Fusion Network for Person Search
Fast Person Search Pipeline

Fig. 4. Milestones of person search studies are steady increase in recent years. As shown in above figure, initially the working on person search started from 2014. Majority of new architectures have developed in recent years because of high demand of community safety applications in person search techniques. After searching on google scholar (October 2019), all papers including journals and conferences are highlighted in figure.

learning, multi-scale feature learning, architecture design, major contribution, loss functions and influence CNNs for person search. Furthermore, the performance of person search methods on famous datasets are carefully evaluated. At last, future insights and further research directions are described. We wish that our article can give researchers with timely reviews and new creativity to facilitate understanding of person search and further bring research on jointly handle detection systems.

1.3. Our Contribution

Our contributions are summarized in following points:

1. Systematic review of person search approaches. We split and summarize the existing retrieval techniques of person search from several respects, including feature learning, multi-scale feature learning, architecture design, loss function and deep metric learning.

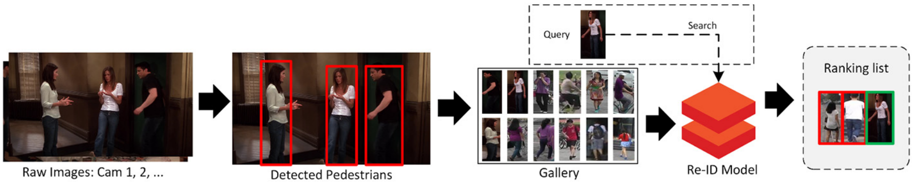

Fig. 5. Simplified view of pedestrian detection and person (re-ID): in gallery, manually cropped images from detected pedestrians' images received from video cameras. Each probe person, system retrieves nearest number of images.

The proposed classification aims to help new researchers with comprehensive and deeper understanding about person search.

2. Comprehensive outlook and analysis of person search performance. Based on our classification of person search approaches, we evaluate and analyze person (re-ID) outcomes of these remarkable methods on existing datasets.
3. We have mentioned several considerable points for developing person search algorithm and giving future insights and directions for researchers.

The outline of the paper is mentioned as: The details and methods of person search (re-ID) are listed in Section 2. Then datasets and the detection performance of small objects are described in Section 3. Finally, we summarized and discussed future research and insights in Section 4.

2. Approaches for person search (re-ID)

One technique is to detect a probe person with gallery of manually cropped person images, little far from real-world applications (see Fig. 5). Jianming et al. [32] presented a technique to understand the spatiotemporal patterns of pedestrians in unlabeled datasets by transferring the visual information from training dataset. The approach did not require knowledge regarding spatial distribution of cameras nor assumption about how persons move in target environment. Jean Paul et al. [33] constructed a generative adversarial network with sparse label smoothing regularization. K-means approached has been used for clustering on training dataset and generated GAN samples for each cluster. Kaiyang et al. [34] developed a lightweight deep CNN architecture by creating a new residual block combined with several convolutional feature streams, each stream detected features in specific manner. Then, unified aggregation network was developed to dynamic fuse multiscale characteristics with input related channel wised weights. Ejaz et al. [35] presented two new layers: a neighborhood layer and subsequent layer in CNN architecture to capture relationship between two images. Hong Xing et al. [36] constructed unsupervised approach through asymmetric deep metric learning technique for person (re-ID) in an unsupervised manner. Certain projection view of each image took from asymmetric clustering on cross-view pedestrian images. The model identifies a share area where viewpoint bias is reduced, and as a result improved matching performance could be achieved. New Deep Association Learning [37] is presented deep learning approach for unsupervised video (re-ID) that extract from unlabeled video from surveillance data. Shuangjie et al. [38] presented ASTPN, joint attentive spatial temporal pooling and similarity measures system in videos that enabled feature extractor to do an ongoing input video sequences, in a way that inter-dependency from same picture can directly impact on computation of each other's representation. Zhedong et al. [39] developed pedestrian alignment network that align pedestrian in bounding box and learns their descriptors. Long et al. [40] developed a framework to track multiple people and constructed scoring function based on convolutional neural network. Yiheng et al. [41] composed a new architecture for video-based person re-identification that depends on two components: Refining Unit and Spatial Clues Integration unit. Hong Xing et al. [42] presented unsupervised loss function to solve asymmetric metric learning problem and pro-posed novel framework DECAMEL that embeds features in end to end learning. Ancong et al. [43] proposed deep model for unsupervised soft multi-label learning. By doing this way, we learn multi-label of each person by matching them with unlabeled person. Shivansh et al. [44] proposed an attention-based architecture for video-based re-identification that calculate frame score where score is higher then what we will evaluate in our experiments. Deqiang et al. [45] presented convolutional network based on two stream fusion: TSF-CNN pick-up spatial and temporal attributes and temporal attention module to allocate a weight of each frame in a sequence. Guangyi et al. [46] developed STAL method that learned attention and spatial temporal dimensions from selective frames for person re-ID and then passed end-to-end network for identification. Another approach in person search (PS) detects all persons and recognize probe person in an image exclusive of proposal and bounding box that is divergent from person re-ID (see Fig. 6). This section primarily focuses on the methods of person search and discusses the importance of literature review. The person search technique is more realistic and applicable in real-world scenarios.

2.1. Feature representation learning

We present several feature learning approaches adopted in person search framework and several papers from person (re-ID) to make better understanding for feature learning. There are two main categories which includes global and local features. Global feature extract features from whole scene without dividing additional cues. Local features divide whole image into part or local features to create a collective representation of each image.

2.1.1. Global features representation learning

Yuyu et al. [58] extracted image features from deep CNN model, attention scheme and bi-LSTM are utilized to encode text information

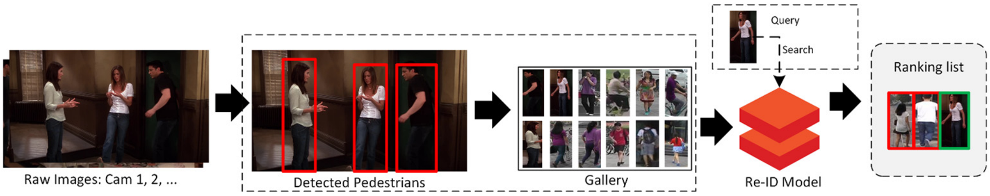

Fig. 6. Simplified view of person search process. Every probe person, we identify possible probe persons in gallery, and compare all possible pairs of the query to identify investigated person.

Jianheng et al. [61] cropped each person image from whole image, global feature map extracted from backbone network. Different RPN [3] is implemented and add person labels to extract deep features from pedestrian whole image. Zhong et al. [25] obtained 512-dimensional feature vector from person image using VGG-16 network. To each node, pairs of features are assigned, and edges are also in same context and a global part is connected in graph to judge whether two input pictures belong to same identity. Graph convolutional networks are applied to combine global and local features [30]. Yichao et al. [22] built a level contextual graph to designed global similarity of investigated gallery sets. Expanded separate characteristics with top result matched context sets, and all these features are exhibited using contextual graph. Cunyuan et al. [62] introduced non-local components that allows model to learn more global features, focus on region in picture where scene is crowded. Zheran et al. [60] proposed multilevel pyramidal feature fusion scheme in the person identification and search. This generates integrated feature maps from top-down to include more comprehensive global and low-level information. The most challenging problem in computer vision is visual query [51,52]: searching a person in image database with the help of free form natural description. Recently, Shuang li et al. [53] studied person search problem with natural language. Recurrent Neural Network (RNN) with Gated Neural Attention (GNA) has been developed to jointly manage person search problem with natural description. Moreover, he collected an extensive dataset with person search description and rich language annotation of person.

2.1.2. Local features representation learning

Local features played a vital and central role in image representation and learning more features based on region aggregated features. In recent years, by combining CNN and local techniques, numerous local features have been proposed for image understanding such as person (re-ID) and person search that jointly operates pedestrian detection and retrieval task in single CNN. With the rapid development of CNN), it becomes the best techniques for image representation, that takes a whole image as input and then train to match its class label. Zhiyuan [50] used 14 patch descriptors which were concatenated to acquire an image-level descriptor. Segmentation algorithm is utilized, and each image is divided into super-pixels. Each high-quality pixel represented as a feature vector based on different features. Color descriptor has been used to extract three dimensional for each pixel in LAB and RGB color space color. Designed re-ID technique where fore person and original image patches divided individually and obtained enriched representations from two separate CNN streams. Two stream networks [55] is designed where foreground features of person and patches of actual image are divided into two sub part to obtain rich representation from stream network. Chuchu et al. [63] used original and foreground image patches into two subnets to enhanced overall accuracy.

The important idea is to combine local part features and full body representation. Jinfu et al. [31] proposed an ELF16: pedestrian image is split into numerous horizontal stripes. Each stripe extracted several texture and color knowledge then combine local features through ensemble. Wei-Hong et al. [26] presented CIF, combined Color Histogram (ColorHist) and Histogram of Oriented Gradient (HOG) to obtain feature representation. Then, every training picture patches and gallery pictures with the combination feature of ColorHist and HOG to acquire better performance. Yan et al. [56] applied convolutional layer (CONV) to merge local features and decrease resolution of lower levels feature maps and spatial features extracted from global average pooling layer.

2.1.3. Multi-scale feature representation learning

The dynamic hierarchical feature fusion (DHFF) [56] is designed to overcome multi-scale matching with multi-level feature fusion. To predict the main hierarchical features of different levels that extract by the backbone network [49], shallow network plays the important part in an attention network. Computed features from each level, weighted by attention values and concatenates to make final fusion feature. Sulan et al. [3] proposed FMT-CNN to solve heterogeneity and correlation tasks, can acquire more robust representation and decrease complexity and training time. More discriminative identity features learn by proposing CLSA [30], structure of feature pyramid and enhanced its representational capacity with semantic alignment learning is employed on single input image. Zheran et al. [60] designed a network that utilized top down and mid-level features to give multi-level feature outputs. Moreover, each feature in mid-level layers work independently in ranking task and scale voting algorithm which votes among these ranking outcomes from different feature levels to obtain ultimate ranking order.

2.1.4. Architecture design

Person search as a pedestrian retrieval task exclusive of proposal and bounding boxes, previous works use the single convolutional neural network architectures [4,27] proposed for pedestrian detection and retrieval as the baseline study. Several researches have attempted to improve the backbone structural design to accomplish superior person (re-ID) features. Convolutional Neural Network (CNN) architectures have played an essential and important part in person (re-ID), because these architectures are helpful for initial feature extraction [47]. The performance of any detector depends on its feature extraction network. Various CNN architectures are developed in the recent years, including VGG [48], ResNet [49], MobileNet [82], DenseNet [83], AlexNet [84], GoogleLeNet [85], Inception [86], etc. VGG16 and ResNet-50 are widely used CNNs for pedestrian detection. Simonyan et al. [48] firstly introduced a heuristic concept by using blocks in deep networks. It is easy to execute repeated block of code through subroutines and loops. The classical convolutional neural network contains a series of layers: i) convolutional layer (CONV) with padding ii) RELU activation function. Each VGG block contains CONV and max pooling. In [48] implement convolutions of 3×3 kernels and max pooling of 2×2 with stride of 2 after each block. In general, VGG network is based on two parts: i) Convolutional and Max Pooling layers (POOL) ii) Fully Connected (FC) layers. The convolutional part connects VGG blocks to achieve success. ResNet-50 introduced a new deep neural network terminology called residual learning. In broader view, deep neural networks are excellent networks for image classification [49]. The researchers go deeper to solve multifaceted problems and achieve high accuracy. By going substantially deeper, training of neural network turns out to be more critical and difficult. Even sometimes this accuracy starts saturating and move down. This problem is solved by Residual learning. The objective of ResNet is to ease training networks by reformulating layers as learning residual block with indication to the inputs layer. ResNet emphasizes on learning residual in place of learning some features. Residual can be easily assumed by subtraction of features learned by input of that layer. Because of this, we easily train neural network and increase more accuracy with degrading previous accuracy.The famous and widely used ResNet-50 [49], an essential implementation is to introduce a non-local layer [62] in the top order of semantic layer after initial layer of convolution layer that used to extract image features. With the combination of RetinaNet [81] with ResNet-50 [49] to developed more powerful extractor. The ResNet blocks and initial convolution are shared between both blocks.

To manage person detection and person re-ID in single architecture design, Xu et al. [11], took an initial step and introduced a person search problem. A sliding window technique has been used to combine with person matching and pedestrian detection. However, this method is limited for handcrafted features. Tong et al. [4,27] developed an optimized framework based on single convolutional network, which jointly handles identification and detection instead of combining detectors and person re-id. An improved hand-crafted feature is proposed [31] to obtain more discriminant and compact features in an identification network. Cunyuan et al. [62] designed an effective combination with Stem CNN [4] and non-local layer to learn discriminant features. He introduced a non-local layer factor of model: the localization rate of pedestrian frame was increased. Zhiyuan et al. [50] proposed supervised and unsupervised re-identification network. In supervised matching technique, every image was symbolized by fourteen patches with k dimensional descriptor. Now, image converted into semantic patch problem, TreeCANN patching algorithm is adopted to calculate the distance between person images. In supervised approach, 14 patches image descriptor concatenated with image descriptor which was used in KLFDA algorithm.

Table 1
Highly influential architectures in top conferences and journals, appearing from - 2014 to 2019 that identifies deep learning for person search task.

Method	Major contributions
OIAM [62]	Designed a non-local module in network that effectively leads the better integration with non-local information. To focus on those areas where pedestrian gather in a scene, pedestrian detector Faster-RCNN modified and implemented structure-aware anchors in detector. Adjust scale and ratio of anchor based on properties of pedestrian box.
PFFN [60]	Proposed PFFN, the goal of network is to fuse multi-level convolutional neural network feature maps in topmost style. Each mid-level feature is independently applying into method of ranking galleries. The PFFN is utilized to generate multi-level output, further, ranking results each mid-level features integrate with scale voting method propose scale to provide an improved ranking order.
IAN [18]	IAN effectively focused on pedestrians and decrease intra distance between feature representations of person. More appropriate and sophisticated region proposals are generated for pedestrians from faster r-cnn in an online manner. Moreover, they thoroughly studied the compatibility of center loss for neural networks, and we discuss the main reasons of center loss and dropout not compatible with each other.
FPSP [61]	Studied person search network with respect to time efficiency and developed faster method called FPSP that runs five times better than existing approaches that are based on faster r-cnn detector [65].
EEPSS [57]	Developed single shot detector based on person search network for person detection and re-ID. They applied triplet loss function to obtain better classification. Furthermore, sequential training of two joint subnets gives more suitable and better performance in re-ID. The shared backbone network builds feature maps which seem to be improved describe person specificities and commonalities.
MCCL [58]	Constructed baseline of person search network with language descriptions. Deep CNN is used to extract visual features of pedestrian and bi-directional LSTM is utilized to encode language descriptions.
Distilled QEEPS [64]	Introduced new approach called knowledge distillation and proposed two separate methods called teacher and student frameworks, for the person detection and for re-ID parts. This technique leads to better model compression without performance dropped.
LCGPS [22]	Introduced a multi-part learning that leads to end-to-end of pedestrian re-ID and multi-part feature learning. The relative model selects informative context of the scene. While considering contextual information, a graph has been used for global similarity between two individuals.
QEEPS [23]	Proposed a novel query guided network that extended the SE-Net block to the investigated channels and spatial features. They defined a novel query similarity subnetwork to learn a query guided re-ID score.
PPCC [21]	Systematically studied the person search problem in videos. Proposed a network that incorporates with identity invariance along a track let and visual similarity. Developed PPCC scheme that substantially increase the reliability of propagation.
IEL [28]	The designed IEL integrates the feature learning process into unlabeled identity information by selectively applying unlabeled identities as enhanced instances. Furthermore, the proposed IEL is generalized and easy to be optimized by using back propagation algorithms.
RCAA [29]	The first investigation about solving person search problem using deep reinforcement learning through conditional decision-making approach. Without proposal computing, proposed model is trained in an end-to-end manner, which may possibly redundant and noisy. They focus temporal contexts and relational spatial into the training process, which guides model to produce more informative like “experience”.

Recently, an efficient I-Net framework [20] constructed with pedestrian proposal networks and shared parameters, structured with Siamese network. The discriminative feature from I-Net fed into two FC layers and extracted further features. Those obtained features are stored in an online dictionary where single positive pair and many negative pairs are generated for computation. Jimin et al. [18] initiated IAN, localized pedestrian and minimize intra-person variation. To predict accurate bounding boxes, region proposal network [19,65] is developed on the top of feature maps obtained from initial part of a network. With an increasing interest in LSTM, a Neural Person Search Machine [54] is developed based on LSTM technique that can preserve spatial knowledge from the spatio-temporal sequences by focusing right region. This model saves contextual information and ignored irrelevant regions. Take query person detail as memory and focused on special regions and emphasis on efficient regions which could be helpful for person search. Di Chen et al. [55] firstly applied faster R-CNN to detect pedestrians and then pass through via re-identification network by modeling original image patches and foreground into subnet to obtain better representation. Inspired by the implementation of person search in real world application, Tong et al. [4] put forward a main contribution which jointly handles person detection and search task. Specially, 256-d features obtained from identification block and then train with OIM [4]. Meanwhile, Jinfu et al. [31] proposed another approach for efficient results. They focused on identification network block; added hand crafted features while CNN block remains same. Later, OIAM [62] introduced two independent CNN blocks. First block was attached with non-local layer to produced image features. Second block sends features vector to FC layer to generate proposal classification and regression, while FMT [3] extended region proposal network [4] by adding anchor re-ids. However, the network improved in overall accuracy and gives valuable results. Yan et al. [56] solve this problem by hierarchical feature fusion approach to learn low features. A shallow network with convolutional and FC layer is created a weight. Each weight will be forward to feature fusion network for classification. Chuchun et al. [63] has put forward a refinement network for person search task. Using detectors, more efficient and reliable bounding boxes are produced for person search network. To generate more accurate bounding boxes, they utilized POI layer which is responsible to crop the detected region from image. Jimin et al. [18] presented a network with feat block which computes distance and other produced location of bounding box. Dingyuan et al. [59] proposed a segmentation masks guided network to avoid negative outcome and focus on background litter part in pedestrian image. The whole network is trained from begin to end manner that considered essential relations among person re-identification, pedestrian detection, and pedestrian segmentation. Thus, further distinct features for detected pedestrians could be retained, which efficiently improves the performance of person search network. As shown in Table 1, we made in-depth study of each person search paper and wrote the significant factors. Xu et al. [30] proposed CLSA technique to addressed multi-scale problem. Finally, they improved architecture of faster-rcnn for consistent pedestrian localization in unconstrained environment, enabling the performance of overall person search task. Di et al. [55] proposed a novel idea to designe separate networks for person detection and re-ID tasks that yields to better performance. One stream developed for foreground region as they focused-on for information and improved feature representation by integrating separate stream from the original image. Panoramic image is entered into object detector that showed bounding boxes and confidence scores as an output. As well as those bounding boxes which have low confidence scores are removed. Hong et al. [24] proposed a new strategy, that makes division of labeled identities enlarge and hardens, division of unlabeled identities soften and minimize intra class distance of feature embedding. In this way, the learning feature of each class is more easy and centralized. Wei-Hong et al. [26] developed CIF framework and considered reformulating the person detection and (re-ID) as a regression goal. Image of probe pedestrian, the given model matched investigated person with the candidates list and judged whether person is pedestrian or not. Meanwhile, discrete Fourier transform approach is employed to increased spatial filtering operation.

2.2. Deep metric learning approach

Deep metric learning is a technique to calculate the distance between two data spaces and decide whether objects are similar or dissimilar. While the approach of deep metric learning is to increase the distance of dissimilar data spaces and reduce the distance between similar objects [56,57] (see Fig. 7). Due to this reason, k-nearest neighbor calculates the distance information and transform into new representation. In recent years, metric learning and deep learning have combine and introduce a new concept called deep metric learning [10]. Yuyu et al. [58] summarized deep metric learning concept for visual objects understanding such as image, speech, video and text. Another most common issue in machine learning is person search. Many CNN networks provide significant success in person search in recent years. Person search aims to identify a person without bounding boxes and proposals. The mostly used benchmarks in person search datasets are CUHK-SYSU and PRW. In [20,24,28,56,58], the authors obtained similarity score based on cosine similarity distance between anchor and positive samples. Yan et al. [22] calculate the distance between pair images using Siamese neural network. Tong et al. [3,27,54,59,60] used Euclidean distance to compute the distance of probe feature and gallery images. Jimin et al. [18] used center loss to decrease intra-class distance for person search

The Euclidean distance defined as

$$d(a,b) = \sqrt{\sum_{i=1}^{k} (a_i - b_i)^2} \tag{1}$$

$$d_M(a_i, a_j) = \sqrt{(a_i - a_j)^T M (a_i - a_j)} \tag{2}$$

dM (a_i, a_j) is distance metric with non-negativity. M must be positive semidefinite and symmetric. All eigenvalues or determinants of M should be zero or positive semidefinite. When we decompose M, as defines

$$M = W^T W \tag{3}$$

$$d_M(a_i, a_j) = \sqrt{(a_i - a_j)^T M (a_i - a_j)}$$

$$d_M(a_i, a_j) = \sqrt{(a_i - a_j)^T W^T W (a_i - a_j)} \tag{4}$$

$$= \|W_{a_i} - W_{a_j}\|_2$$

In Eq. (4), W is a linear transformation property. Now Euclidean distance is transforming into Mahalanobis distance for two samples.

Cosine similarity metric defined as

$$S_{cos} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}} \tag{5}$$

2.2.1. Novel loss functions

Loss function played a fundamental role in person search task because they are calculating a difference of actual and predicted value. A number of famous loss function have been proposed in top venues in recent years including OIAM [62], Center loss [18], Multi-Metric [56], MCCL [58], Proxy Triplet [63], OLP [20], IOIM [24], IEL [28] and OIM [4] are designed for person search. Moreover, OIM [4] is one of the famous loss functions used in person search and person re-ID task. Especially, these loss functions are developed to solve person search problem. (For more description, see Table 2.)

3. Datasets and performance evaluation

This section refers to the frequently used datasets for the evaluation to achieve person search task with good performance. Huge number of benchmarking datasets has been published to achieve person re-identification and person search task. Benchmarking datasets are listed in (see Table 3). The most common evaluated benchmarks are CUHK-SYSU and PRW. Several person re-id datasets are small in term of data to tackle person re-id issue.

3.1. Datasets

Constructing person re-id dataset is a critical and security risk task as it requires permission from individual person under uncontrolled environment. Despite its importance, person search has gained significant demand in the market. It is mainly because of easier process from person re- identification method. First, large-scale of real-time data is needed to train a deep learning model. Some datasets are (see Table. 3) are small, especially PRW and VIPER have less than 1000 IDs. Recently, Tong et al. [4] released a new dataset designed for person search benchmarking. This dataset contains approximately 9000 IDs and almost 100,000 bounding boxes which provides good amount of data. The scientific community needs more large datasets to achieve good accuracy. Second, the detection of candidate person in an image is always challenging such as low resolution, occlusion, and misalignment etc. issues. As a result, the chances of false detection may be increased, which is inevitable for

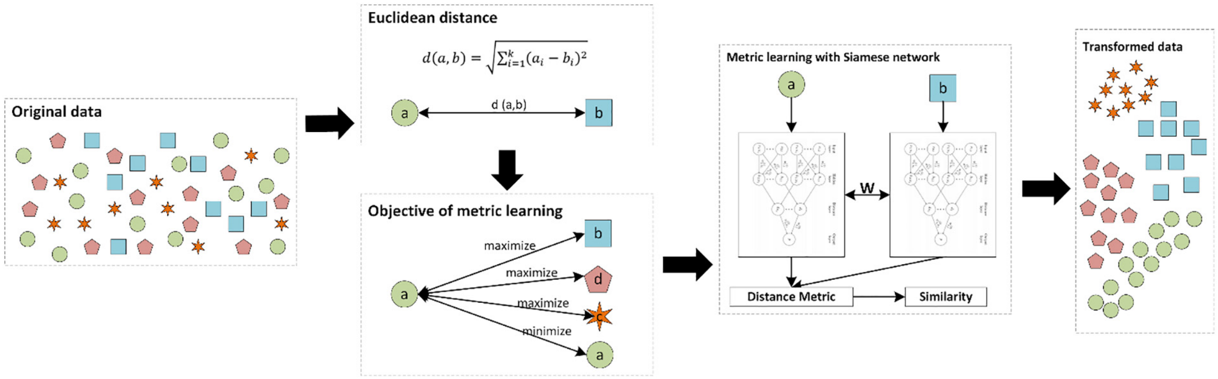

Fig. 7. Simplified view of deep metric learning: Applying Euclidean distance on data space, metric learning is decreasing the distance between two similar data spaces and increase the distance between dissimilar data spaces. This can be achieved using Siamese neural network. Finally, data space is transformed into cluster shape.

Table 2
An overview of novel loss functions.

Loss function	Description	Published
OIAM [62]	An online instance aggregation matching put forward to localize persons with many identities, it does not only depend on intra-class distance.	Neurocomputing, 2019
Center Loss [18]	The aim of center loss is to minimize intra-class distance by combining same features in each class. It will be helpful in feature compactness without the need of positive and negative samples.	Pattern Recognition, 2019
Multi-Metric [56]	More convenient way to minimize a loss by measuring only the features of labeled and unlabeled IDs. All the features are based on hard triplet mining method.	ICIP, 2019
MCCL [58]	A mutual connected loss function emphasizes on identity level information. It introduced identification information into image and language description. It focused on probability of cross modal classification of same identities to become more similar	ICASSP, 2019
Proxy Triplet [63]	A proxy approach contains anchor, positive and negative triplets. In initial stage of training, each proxy contains zeros in proxy table. During backpropagation, proxy table updated with the sample values.	ICCV, 2019
OLP [20]	A dictionary has been designed where the feature of each person with its background proposal and labels are stored. The stored features directly correspond to mini batch size. When loss function obtains more features, it will be replaced from previous features.	ACCV, 2018
IOIM [24]	Inspired by OIM [4], the objective is to make unlabeled identities nearer by minimizing the probability of unlabeled identities. While distribution of labeled identities made it much harder.	ICASSP, 2018
IEL [28]	Instance enhancing loss function is focused on unlabeled identities to learn deep discriminative identities. An enhanced instance produced for unlabeled information and easy to optimize using backpropagation algorithms.	ICIP, 2018
OIM [4]	A non-parametric loss function directly learns from features without a need of huge classifier matrix. Furthermore, sub-sampling approach for labeled and unlabeled identities to decrease computation time.	CVPR, 2017

Table 3
Statistics of commonly used datasets in person search papers for person re-ID.

Dataset	# ID	# BBox	# Cam	# Lab	Evaluation criteria
CUHK-SYSU	8432	99,809	N/a	Hand	CMC & mAP
Person re-ID in wild (PRW)	932	34,304	6	Hand	mAP
Motion analysis and re-identification set (MARS)	1261	1,191,003	6	Hand	CMC & mAP
Cast search in movies (CSM)	1218	700 K	N/a	Hand	mAP
Visual person detection made reliable (VIPeR)	632	1264	2	Hand	CMC
Person re-ID (PRID450s)	450	900	2	Hand	CMC
CUHK01	971	3884	2	Hand	CMC
Pedestrian attribute (PETA)	8705	N/a	N/a	N/a	ikSVM, MR (Fg, Fr)
EPFL	30	N/a	4	N/a	N/a
CAMPUS-human	74	N/a	N/a	N/a	CMC

gallery. Therefore, researchers may construct new techniques to avoid misalignment and false detection. Third, more accurate and real-time data are used during dataset collection. Tong et al. [4] used Street snap and movies data which is little bit high resolution of actual CCTV camera images. When it comes to person re-identification, widespread cameras in city will be helpful to make accurate and robust deep learning models.

3.2. Evaluation metrics

Evaluating person search and re-id methods, the Cumulative Matching Characteristics (CMC) is employed. CMC presents probability of query image that appears in candidate list with sized difference. Considering right ones, only first image is counted in CMC calculation. Therefore, CMC is precise technique when factual data of a query exists. This criterion follows when researchers care more about returning ground truth (GT) in top position in ranking list.

While multiple ground truth exists in gallery, Mean Average Precision (mAP) is best way for evaluation in this scenario. mAP is used when two cameras spotting same GT but have uncommon retrieval recall capability. In this moment, mAP works fine instead of CMC because CMC have less discriminative ability. In addition to person search task, a candidate box considered positive if GT is greater than 0.5. Later in, many research papers [62,60,3,56,57,58,63,64] reported mAP result with multiple ground truth. CUHK-SYSU contains 18,184 images, 96,143 pedestrians by ignoring background pedestrians whose height are >50 pixels. The persons appeared in different cameras, resulting, labeled 8432 identities. PRW (Person-Re-identification in the Wild) is an extended version of Market 1501 dataset. They acquire original videos from university campus which contain six cameras, among five were high quality (1080 × 1920 HD) and 1 is low quality (576 × 720 SD). It contains 11,816 frames, 43,110 pedestrians bounding boxes and 34,304 pedestrians were annotated with unique identity with 1 to 932. MARS, a more extended version of Market-1501 especially designed for video-based person re-id. During data collection, six cameras placed in Tsinghua University. Five cameras were of high resolution (1080 × 1920 HD) and one camera with a low resolution (640 × 480 SD). It contains 1262 IDs, 1,067,516 bounding boxes and 3248 distractors. Each pedestrian was captured from two different cameras. CSM, a diverse dataset consists of 127 K track-lets of 1218 identities from 192 movies. All identities are manually annotated, contains 11 M instances and came with reference portrait. This benchmark is quite challenging due to illumination, makeup, pose, clothing and age factor. VIPeR, introduced a viewpoint invariant dataset of pedestrians. It contains 632 identities of pedestrian image pairs with different pose and illumination changes. Data is collected from two cameras, contains 1264 images and all the images were manually annotated with crop size (128 × 48). Person Re-ID (PRID450s) a more realistic dataset builds on PRID 2011 and arranged by image pairs. The dataset consists of 450 single shots of walking humans captured from two disjoint camera views. From the perspectives, the pedestrian images with resolution of (720 × 576 pixels). Each image of pedestrian was annotated with size (100–150 pixels). CUHK01 Well-known publicly available dataset is collected by W. Li et al. named CUHK01. It contains 971 pedestrian images of two disjoint cameras in college campus. All the images are manually annotated and contain multi-shot images with size (160 × 60). PETA (Pedestrian Attribute) dataset is the composition of true pedestrian images taken 7% from VIPER, 5% of 3DPES, 6% from CAVIR4REID, 24% CUHK, 7% GRID, 2% i-LIDS, 5% MIT, 6% PRID, 1% SARC3D and 37% from Town Centre. It contains 19,000 images, 8705 person images. Each image is annotated with four multi-class attributes. In this dataset, variation such as camera angle, viewpoints, illumination, indoor and outdoor scenes are included. In addition, many person images have low resolution. These factors make person re-identification more challenging. CAMPUS-Humam, Yuanlu Xu et al. constructed a dataset for person-re-id in surveillance applications. There are 74 identities, with IDs and location included. It contains 370 reference images with 175 pixels in height. Moreover, consecutive images of each person taken from diverse pose, views, occlusion and conjunction. EPFL, along with the above-mentioned person search and re-identification datasets, original dataset has been proposed to track person in multiple views. The dataset included 30 IDs, 294 bounding boxes and four cameras.

3.3. Performance analysis

We have thoroughly summarized the methods of person searching in an image. Based on our person search algorithm, we show the search results of these state-of-the-art algorithms on CHUK-SYSU and PRW datasets in Tables 4 and 5 respectively. By comparing Tables 4 and 5, we can see that the person search task is difficult in PRW. This is maybe because the number of person is higher; occlusion and variation are much greater than CHUK-SYSU dataset. Compared with CHUK-SYSU, the several objects in scene are not clear, which effects on overall results.

As shown in Table 5, the searching performance of person is the best one in CHUK-SYSU dataset. This dataset specially designed for person search task. LRPS [63] improves 88.5 to 94.2 on top-1 rank, and also enhances 87.2 to 93.0 mAP. Moreover, utilizing ROI transform layer has been added to produce more accurate bounding boxes and cropping image from whole scene.

4. Conclusion and future directions

Person search is a new paradigm which is practically applicable in large communities without investing huge time on manually cropped images of persons. In this review paper, recent works of person search are discussed. Firstly, we have discussed person search history and trend in recent years. Secondly, detailed process of person search and person re-identification with literature review is discussed. Thirdly, we reviewed CNN existing architectures used in person search to achieve it. Then, highly cited person search architectural were described with diagram illustration. Finally, we also gave the metric learning solutions for person search. Prospective of video surveillance systems, person search has increased a lot of consideration in computer vision industries. Therefore, we emphasize on developed work on person search

Our review studied of person search algorithms and datasets help us to explore re-id algorithm that jointly manages person detection and re-ID task. It needs to be done in future to make more it robust and accurate. To conclude our person search review study, we discuss several ideas from this study, as well as proposed research directions and suggestions for re-ID researchers that might be helpful to develop effective and better algorithms

Table 4
Comparative results on CHUK-SYSU test set. Mostly methods belong to ResNet-50, ResNet-101, and only one of VGG-16 as the backbone of network. "Large Scale" in gallery size denoted complete dataset used for testing.

Method	Backbone	Gallery size	top-1(%)	top-5(%)	mAP (%)
OIAM [62]	ResNet-50	100	77.86	90.56	76.98
		LS	78.28	90.41	77.63
PFFN [60]	ResNet-101	100	89.8	–	84.5
FMT [3]	ResNet-50	100	79.83	90.90	77.15
SMGPS [59]		100	86.5	–	86.3
IAN [18]	ResNet-101	Large Scale	80.45	–	77.23
DHFF [56]	ResNet-50	100	91.7	–	90.2
FPSP [61]		100	89.87	–	86.99
EEPSS [57]		100	80.5	–	79.4
LRPS [63]		100	94.2	–	93.0
Distilled QEEPS [64]		Large Scale	85.5	–	85.0
LCGPS [22]		100	86.5	–	84.1
QEEPS [23]		100	89.1	–	88.9
I-Net [20]	VGG-16	Large Scale	81.5	–	79.5
MGTS [55]	ResNet-50	100	90.7	–	89.1
IOIM [24]		100	79.90	–	79.78
IEL [28]		100	79.66	–	79.43
RCAA [29]		Large Scale	81.3	–	79.3
CLSA [30]		100	88.5	–	87.2
NPSM [54]		100	81.2	–	77.9
OIM [4]		Large Scale	78.7	–	75.5
ELF16 [31]		Large Scale	80.6	–	77.8

Table 5
Comparative results on PRW test set. All methods utilized ResNet-50 and ResNet-101 as the backbone of network. "Large Scale" in gallery size denoted complete dataset used for testing.

Method	Backbone	Gallery size	top-1(%)	top-5(%)	mAP (%)
OIAM [62]	ResNet-50	200	69.85	87.70	51.02
PFFN [60]	ResNet-101	4000	73.9	–	34.3
IAN [18]		Large Scale	61.85	–	23.00
DHFF [56]	ResNet-50	Large Scale	70.1	–	41.1
FPSP [61]		Large Scale	70.58	–	44.45
EEPSS [57]		Large Scale	47.0	–	25.2
LRPS [63]		Large Scale	70.2	–	42.9
Distilled QEEPS [64]		Mini	80.0	–	39.7
LCGPS [22]		Large Scale	73.6	–	33.4
QEEPS [23]		Large Scale	80.0	–	39.1
MGTS [55]		Large Scale	72.1	–	32.6
IOIM [24]		Large Scale	63.10	–	21.00
IEL [28]		Large Scale	69.47	–	24.26
CLSA [30]		Large Scale	65.0	–	38.7
NPSM [54]		100	53.1	–	24.2

We considered that ResNet-50 given the foremost feature extraction performance including all CNN models, with VGG-16 and MobileNet far away. Again, the performance is surely due to generalizable features and learning enough features to move on further step. While adapting new architectures such as ResNet-101 will naturally give better performance. The primary reason for its success is the residual blocks with stacking convolution layers. There are two key points from this observation: Mitigate vanishing gradient problem to follow alternate way of gradient flow and identity function to ensure that higher layer perform good [49]. Recent advancement in facture learning would bring more promising research direction to learn local patches from images and then pass to network using existing scheme. Faster RCNN [65] is a best region proposal candidate to generate local patches on image and trained end to end way. There are several new architectures [4,18,56] that manages pedestrian detection and re-ID in one process. Consequently, recent advances in person search networks such as OIM, IAN or DHFF specially designed with convolutions would be more relevant

To consider real-world scenario, the annotation and bounding boxes of each person with unique identity is not present [27]. We can potentially make new architectures free from bounding boxes [4,27,56] to fulfill surveillance needs. An immediate way is to develop new benchmark datasets and more progressive approaches that automatically detect person from image. Existing pedestrian detectors [65] may cause misalignment and misdetection under various conditions. Another more promising area in person search task; is automatic pedestrian detection from whole scene of images. While issues like automatic pedestrian detection may cause misalignment and missed pedestrian in first step. We can solve this problem by proposing new detectors and passing an image from multiple detectors and trained detectors in more efficient way

While the existing person search algorithms directly take pedestrian image for person detection. Low resolution and noisy images may decrease system performance. We can improve it by adding new network [66] before detector which minimizes the noise within an image. In addition to implement this strategy, we restore more blurred images in original image. A sequence of low-resolution images obtains from CCTV camera, we could generate more images by borrowing the idea from [67] available images. Furthermore, these techniques increase overall accuracy of person search task. An interesting recent work which will be helpful to make person search system more useful is to add contextual information and behavior of each person gives more context in surveillance system. Integrating the idea of natural language description [53] would be promising future research direction to identify criminals, weapon equipped person etc. will be ultimate destination of person search task

In general, for re-id and person search task, the special and major attribute is human walking patterns and gaits are critical [68]. This walking style [69] gives us information to identify a person in crowded environment. While there are some research work in person re-identification [70,71], more work is needed in person task, especially implementing gait with appearance and feature learning could be good starting point for further research. Other future research area includes multi-modal data to tackle appearance problem. For example, different companies have strict policy to wear uniform for all workers. In such scenarios, re-id algorithms are failed

To sum up, we conclude with several ideas on the current person search task and performance evaluation for re-id algorithms. We believe that researchers should focus on developing new techniques and not just look on performance improvement factors under same circumstances. CUHK-SYSU [4] is small benchmark dataset and collected images from several television show. To this end, we must emphasis on creating large and real-world perspectives where system will be deployed

Declaration of Competing Interest

None.

References

[1] Z. Liang, H. Zhang, S. Sun, M. Chandraker, Y. Yang, Q. Tian, Person re-identification in the wild, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2017) 1367–1376.

[2] L. Zheng, Y. Yang, A.G. Hauptmann, Person re-identification: Past, present and future, arXiv Preprint, 2016arXiv:1610.02984.

[3] S. Zhai, S. Liu, X. Wang, J. Tang, FMT: fusing multi-task convolutional neural network for person search, Multimed. Tools Appl. (2019) 1–12.

[4] T. Xiao, S. Li, B. Wang, L. Lin, Xi Wang, Joint detection and identification feature learning for person search, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, pp. 3415–3424.

[5] X. Chu, W. Ouyang, H. Li, X. Wang, Structured feature learning for pose estimation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 4715–4723.

[6] Wei Li, Rui Zhao, Tong Xiao, Xiaogang Wang, Deepreid: Deep filter pairing neural network for person re-identification, Proceedings of the IEEE conference on computer vision and pattern recognition 2014, pp. 152–159.

[7] Shengcai Liao, Yang Hu, Xiangyu Zhu, Z.Li. Stan, Person re-identification by local maximal occurrence representation and metric learning, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015, pp. 2197–2206.

[8] Sakrapee Paisitkriangkrai, Chunhua Shen, Anton Van Den Hengel, Learning to rank in person re-identification with metric ensembles, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015, pp. 1846–1855.

[9] Tong Xiao, Hongsheng Li, Wanli Ouyang, Xiaogang Wang, Learning deep feature representations with domain guided dropout for person re-identification, Proceedings of the IEEE conference on computer vision and pattern recognition 2016, pp. 1249–1258.

[10] Liang Zheng, Liyue Shen, Tian Lu, Shengjin Wang, Jingdong Wang, Qi Tian, Scalable person re-identification: A benchmark, Proceedings of the IEEE international conference on computer vision 2015, pp. 1116–1124.

[11] X. Yuanlu, M. Bingpeng, R. Huang, L. Lin, Person Search In A Scene By Jointly Modeling People Commonness And Person Uniqueness, Proceedings Of The 22nd ACM International Conference On Multimedia November 03–07, 2014, pp. 937–940.

[12] S. Gong, M. Cristani, C.C. Loy, M.T. Hospedales, Person Re-Identification (Advances in Computer Vision and Pattern Recognition), 2014 301–313.

[13] Qingming Leng, Mang Ye, Qi Tian, A survey of open-world person re-identification, IEEE Transactions on Circuits and Systems for Video Technology 30 (4) (April 2020) 1092–1108.

[14] T. D'Orazio, G. Cicirelli, People re-identification and tracking from multiple cameras: a review, 2012 19th IEEE International Conference on Image Processing, IEEE 2012, pp. 1601–1604.

[15] A. Bedagkar-Gala, S.K. Shah, A survey of approaches and trends in person re-identification, Image Vis. Comput. 32 (4) (2014) 270–286.

[16] S. Gong, M. Cristani, S. Yan, C.C. Loy, Person re-identification, 1, Springer, 2014.

[17] R. Satta, Appearance descriptors for person re-identification: a comprehensive review, arXiv preprint, 2013arXiv:1307.5748.

[18] Jimin Xiao, Yanchun Xie, Tammam Tillo, Kaizhu Huang, Yunchao Wei, Jiashi Feng, Ian: the individual aggregation network for person search, Pattern Recogn. 87 (2019) 332–340.

[19] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2014, pp. 580–587.

[20] Z. He, L. Zhang, End-to-end detection and re-identification integrated net for person search, Asian Conference on Computer Vision, Springer 2018, pp. 349–364.

[21] Q. Huang, W. Liu, D. Lin, Person search in videos with one portrait through visual and temporal links, Proceedings of the European Conference on Computer Vision 2018, pp. 425–441.

[22] Y. Yan, Q. Zhang, B. Ni, W. Zhang, M. Xu, X. Yang, Learning context graph for person search, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2019) 2158–2167.

[23] B. Munjal, S. Amin, F. Tombari, F. Galasso, Query-guided end-to-end person search, CVPR, 2019.

[24] H. Liu, W. Shi, W. Huang, Q. Guan, A Discriminatively Learned Feature Embedding Based on Multi-Loss Fusion for Person Search, IEEE International Conference on Acoustics, Speech and Signal Processing, 2018 1668–1672.

[25] Z. Ji, S. Li, Y. Pang, Fusion-attention network for person search with free-form natural language, Pattern Recogn. Lett. 116 (2018) 205–211.

[26] W.H. Li, Y. Mao, A. Wu, W.S. Zheng, Correlation based identity filter: An efficient framework for person search, International Conference on Image and Graphics (2017) 250–261.

[27] T. Xiao, S. Li, B. Wang, L. Lin, X. Wang, End-to-end deep learning for person search, arXiv Preprint, 2016 arXiv:1604.01850 , 2:2.

[28] W. Shi, H. Liu, F. Meng, W. Huang, Instance enhancing loss: deep identity-sensitive feature embedding for person search, IEEE International Conference on Image Processing, Pages 4108–4112 (2018).

[29] X. Chang, P.Y. Huang, Y.D. Shen, X. Liang, Yi Yang, A.G. Hauptmann, Rcaa: Relational context-aware agents for person search, Proceedings of the European Conference on Computer Vision (2018) 84–100.

[30] X. Lan, X. Zhu, S. Gong, Person search by multi-scale matching, Proceedings of the European Conference on Computer Vision (2018) 536–552.

[31] Jinfu Yang, Meijie Wang, Mingai Li, Jingling Zhang, Enhanced deep feature representation for person search, CCF Chinese Conference on Computer Vision, Springer 2017, pp. 315–327.

[32] J. Lv, W. Chen, Q. Li, C. Yang, Unsupervised cross- dataset person re-identification by transfer learning of spatial-temporal patterns, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 7948–7956.

[33] Jean-Paul Ainam, Ke Qin, Guisong Liu, Guangchun Luo, Sparse la- bel smoothing regularization for person re-identification, IEEE Access 7 (2019) 27899–27910.

[34] K. Zhou, Y. Yang, A. Cavallaro, T. Xiang, Omni-scale feature learning for person re-identification, ICCV, 2019.

[35] E. Ahmed, M. Jones, T.K. Marks, An improved deep learning architecture for person re-identification, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2015) 3908–3916.

[36] Hong-Xing Yu, Ancong Wu, Wei-Shi Zheng, Cross-view asymmetric metric learning for unsupervised person re-identification, Proceedings of the IEEE International Conference on Computer Vision 2017, pp. 994–1002.

[37] Yanbei Chen, Xiatian Zhu, Shaogang Gong, Deep association learning for unsupervised video person re-identification, arXiv preprint, 2018arXiv:1808.07301.

[38] S. Xu, Y. C, K. Gu, Y. Yang, S. Chang, P. Zhou, Jointly attentive spatial-temporal pooling networks for video-based person re-identification, Proceedings of the IEEE International Conference on Computer Vision 2017, pp. 4733–4742.

[39] Z. Zheng, L. Zheng, Y. Yang, Pedestrian alignment network for large-scale person re-identification, IEEE Transactions on Circuits and Systems for Video Technology, 2018.

[40] C. Long, A. Haizhou, Z. Zijie, S. Chong, Real-time multiple people tracking with deeply learned candidate selection and person re-identification, ICME, 5, 2018, p. 8.

[41] Yiheng Liu, Zhenxun Yuan, Wengang Zhou, Houqiang Li, Spatial and temporal mutual promotion for video-based person re-identification, arXiv preprint, 2018arXiv:1812.10305.

[42] Hong-Xing Yu, Ancong Wu, Wei-Shi Zheng, Unsupervised person re- identification by deep asymmetric metric embedding, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018.

[43] Hong-Xing Yu, Wei-Shi Zheng, Ancong Wu, Xiaowei Guo, Shaogang Gong, Jian-Huang Lai, Unsupervised person re-identification by soft multil- abel learning, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019, pp. 2148–2157.

[44] Shivansh Rao, Tanzila Rahman, Mrigank Rochan, Wang Yang, Video- based person re-identification using spatial-temporal attention networks, arXiv preprint, 2018arXiv:1810.11261.

[45] D. Ouyang, Y. Zhang, J. Shao, Video-based person re- identification via spatio-temporal attentional and two-stream fusion convolutional networks, Pattern Recogn. Lett. 117 (2019) 153–160.

[46] Guangyi Chen, Jiwen Lu, Ming Yang, Jie Zhou, Spatial-temporal attention-aware learning for video-based person re-identification, IEEE Trans. Image Process. 28 (9) (Sept. 2019) 4192–4205.

[47] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, ImageNet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.

[48] Karen Simonyan, Andrew Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint, 2014arXiv:1409.1556.

[49] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 770–778.

[50] Zhiyuan Shi, Timothy M. Hospedales, Tao Xiang, Transferring a semantic representation for person re-identification and search, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015, pp. 4184–4193.

[51] Mengye Ren, Ryan Kiros, Richard Zemel, Image question answering: a visual semantic embedding model and a new dataset, Proc. Advances in Neural Inf. Process. Syst 1 (2) (2015) 5.

[52] Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, Rob Fergus, Simple baseline for visual question answering, arXiv preprint, 2015arXiv:1512.02167.
[53] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, X. Wang, Person search with natural language description, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2017) 1970–1979.
[54] H. Liu, J. Feng, Z. Jie, K. Jayashree, B. Zhao, M. Qi, J. Jiang, S. Yan, Neural person search machines, Proceedings of the IEEE International Conference on Computer Vision (2017) 493–501.
[55] C. Di, S. Zhang, W. Ouyang, J. Yang, Y. Tai, Person Search Via A Mask-Guided Two-Stream CNN Model, Proceedings of the European Conference on Computer Vision 2018, pp. 734–750.
[56] Yan Lu, Zheran Hong, Bin Liu, Weihai Li, Yu. Nenghai, Dhff: Robust multi-scale person search by dynamic hierarchical feature fusion, 2019 IEEE International Conference on Image Processing (ICIP), IEEE 2019, pp. 3935–3939.
[57] Angelique Loesch, Jaonary Rabarisoa, Romaric Audigier, End-To-End person search sequentially trained on aggregated dataset, 2019 IEEE International Conference on Image Processing (ICIP), IEEE 2019, pp. 4574–4578.
[58] Yuyu Wang, Chunjuan Bo, Wang Dong, Shuang Wang, Yunwei Qi, Huchuan Lu, Language person search with mutually connected classification loss, ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE 2019, pp. 2057–2061.
[59] Dingyuan Zheng, Jimin Xiao, Kaizhu Huang, Yao Zhao, Segmentation mask guided end-to-end person search, arXiv preprint, 2019arXiv:1908.10179.
[60] Zheran Hong, Bin Liu, Yan Lu, Guojun Yin, Yu Nenghai, Scale Voting with Pyramidal Feature Fusion Network for Person Search, IEEE Access, 2019.
[61] Jianheng Li, Fuhang Liang, Yuanxun Li, Wei-Shi Zheng, Fast Person Search Pipeline, 2019 IEEE International Conference on Multimedia and Expo (ICME), IEEE 2019, pp. 1114–1119.
[62] Cunyuan Gao, Rui Yao, Jiaqi Zhao, Yong Zhou, Fuyuan Hu, Leida Li, Structure-aware person search with self-attention and online instance aggregation matching, Neurocomputing 369 (2019) 29–38.
[63] Chuchu Han, Jiacheng Ye, Yunshan Zhong, Xin Tan, Chi Zhang, Changxin Gao, Nong Sang, Re-ID driven localization refinement for person search, ICCV (2019) 9814–9823https://ieeexplore.ieee.org/document/9010724.
[64] Bharti Munjal, Fabio Galasso, Sikandar Amin, Knowledge distillation for end-to-end person search, BMVC (2019)https://bmvc2019.org/wp-content/uploads/papers/0198-paper.pdf.
[65] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, Advances in Neural Information Processing Systems 2015, pp. 91–99.
[66] Lingxiao Wang, Yali Li, Shengjin Wang, DeepDeblur: fast one-step blurry face images restoration, arXiv preprint, 2017arXiv:1711.09515.
[67] Gaohua Liao, Quanguo Lu, Xunxiang Li, Research on fast super-resolution image reconstruction base on image sequence, 2008 9th International Conference on Computer-Aided Industrial Design and Conceptual Design, IEEE 2008, pp. 680–684.
[68] J.E. Cutting, L.T. Kozlowski, Recognizing friends by their walk: gait perception without familiarity cues, Bull. Psychon. Soc. 9 (5) (1977) 353–356.
[69] E.-S.M. El-Alfy, A.G. Binsaadoon, Automated gait-based gender identification using fuzzy local binary patterns with tuned parameters, J. Ambient. Intell. Humaniz. Comput. 10 (7) (2019) 2495–2504.
[70] Cassandra Carley, Ergys Ristani, Carlo Tomasi, Person Re-Identification from Gait using an Autocorrelation Network, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019 , pp. 0–0.
[71] Apurva Bedagkar-Gala, Shishir K. Shah, Gait-assisted person re-identification in wide area surveillance, Asian Conference on Computer Vision, Springer, Cham 2014, pp. 633–649.
[72] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Fu Cheng-Yang, Alexander C. Berg, Ssd: Single shot multibox detector, European Conference on Computer Vision, Springer, Cham 2016, pp. 21–37.
[73] Mengran Gou, Ziyan Wu, Angels Rates-Borras, Octavia Camps, Richard J. Radke, A systematic evaluation and benchmark for person re-identification: features, metrics, and datasets, IEEE Trans. Pattern Anal. Mach. Intell. 41 (3) (2018) 523–536.
[74] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, Steven C.H. Hoi, Deep learning for person re-identification: a survey and outlook, arXiv Preprint, 2020arXiv:2001.04193.
[75] Bahram Lavi, Mehdi Fatan Serj, Ihsan Ullah, Survey on deep learning techniques for person re-identification task, arXiv preprint, 2018arXiv:1807.05284.
[76] Kejun Wang, Haolin Wang, Meichen Liu, Xianglei Xing, Tian Han, Survey on person re-identification based on deep learning, CAAI Transactions on Intelligence Technology 3 (4) (2018) 219–227.
[77] Athira Nambiar, Alexandre Bernardino, Jacinto C. Nascimento, Gait-based person re-identification: a survey, ACM Computing Surveys (CSUR) 52 (2) (2019) 1–34.
[78] Zheng Wang, Zhixiang Wang, Yinqiang Zheng, Yang Wu, Wenjun Zeng, Shin'ichi Satoh, Beyond Intra-modality: A Survey of Heterogeneous Person Re-identification, IJCAI, 2020.
[79] Mohammad Ali Saghafi, Aini Hussain, Halimah Badioze Zaman, Mohamad Hanif Md Saad, Review of person re-identification techniques, IET Computer Vision 8 (6) (2014) 455–474.
[80] Di Wu, Si-Jia Zheng, Xiao-Ping Zhang, Chang-An Yuan, Fei Cheng, Zhao Yang, Yong-Jun Lin, Zhong-Qiu Zhao, Yong-Li Jiang, De-Shuang Huang, Deep learning-based methods for person re-identification: a comprehensive review, Neurocomputing 337 (14 April 2019) 354–371.
[81] T.-Y. Lin, P. Goyal, R.B. Girshick, K. He, P. Dollar, Focal loss for dense object detection, IEEE TPAMI, 2018.
[82] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint, 2017arXiv:1704.04861.
[83] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, Kilian Q. Weinberger, Densely connected convolutional networks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, pp. 4700–4708.
[84] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in Neural Information Processing Systems 2012, pp. 1097–1105.
[85] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, Going deeper with convolutions, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015, pp. 1–9.
[86] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, Zbigniew Wojna, Rethinking the inception architecture for computer vision, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 2818–2826.