Person Search Challenges and Solutions: A Survey

Xiangtan Lin\textsuperscript{1}, Pengzhen Ren\textsuperscript{2}, Yun Xiao\textsuperscript{2}, Xiaojun Chang\textsuperscript{1*}, Alex Hauptmann\textsuperscript{3}

\textsuperscript{1}Monash University
\textsuperscript{2}Northwest University
\textsuperscript{3}Carnegie Mellon University

\{xiangtan.lin, cxj273, alex.hauptmann\} @gmail.com, pzhren@foxmail.com, yxiao@nwu.edu.cn

Abstract

Person search has drawn increasing attention due to its real-world applications and research significance. Person search aims to find a probe person in a gallery of scene images with a wide range of applications, such as criminals search, multi-camera tracking, missing person search, etc. Early person search works focused on image-based person search, which uses person image as the search query. Text-based person search is another major person search category that uses free-form natural language as the search query. Person search faces more challenges than person re-id problem. Unlike the person re-id setting where the cropped person images are provided, and the primary challenge is just to bring the query-person gap. Person search needs to deal with an additional detection challenge so that the detected person can be used for the downstream identification task. Such detection results may be inconsistent with the identification task (Figure 2). Similarly, text-based person search is also more challenging than the traditional text-image matching problem [Li et al., 2017b] as it needs to learn discriminative features first before the text-person matching.

Person search is fast-evolving, and existing person search methods are diverse and complex. Researchers may leverage the rich knowledge concerning object detection, person re-id, and text-image matching separately. Systematic surveys

1 Introduction

Person search [Xu et al., 2014] aims to find a probe person in a gallery of scene images. Historically, person search was an extended form of person re-identification (re-id) problem [Liu et al., 2020b; Liu et al., 2020a; Li et al., 2019b; Cheng et al., 2018; Liu et al., 2018b; Cheng et al., 2017; Liu et al., 2017b]. Therefore, early researches on person search focused on an image-based setting, which uses person image as the search query [Xiao et al., 2017; Liu et al., 2017a; Chang et al., 2018; Gao et al., 2019; Xiao et al., 2019]. Meanwhile, research in text-based person search [Li et al., 2017b; Wang et al., 2020b] has made significant advances in the past few years. Text-based person search is handy when a probe image is unavailable but free-form natural language. The two types of person search are illustrated in Figure 1.

Figure 1: The general frameworks of person search. (a) Image-based person search in which person image is available as the search query against a gallery of images. Image-based person search involves two sub-tasks, person detection and person identification. (b) Text-based person search in which search query is free form natural language. A general text-based person search framework typically learns text feature through an RNN variant network and then align text features with visual elements from the detection network to identify the person in the target images.

*Corresponding Author
Concerning person search bring more values to the community. Especially, as far as we know, there is no existing survey covering the text-based person search. [Islam, 2020] surveyed works on image-based person search and neglected the text-based person search. Furthermore, [Islam, 2020] didn’t discuss the joint challenge of person detection and identification, especially the detection-identification inconsistency challenge as illustrated in Figure 2. Therefore, we survey works beyond image-based person search and provide a systematic review of the diverse person search solutions. We summarise the main differences between the previous survey [Islam, 2020] and ours in Table 1.

In this survey, we aim to provide a cohesive analysis of the recent person search works so that the rationals behind the ideas can be grasped to inspire new ideas. Specifically, we surveyed recently published and pre-print person search papers from top conference venues and journals. We analyse methods from the perspective of challenges and solutions and summarise evaluation results accordingly. At the end of the paper, we provide insights on promising future research directions. In summary, the main contributions of this survey are:

- In addition to image-based person search, we cover text-based person search which was neglected in the previous person search survey.
- We analyse person search methods from the perspective of challenges and solutions to inspire new ideas.
- We summarise and analyse existing methods’ performance and provide insights on promising future research directions.

### 2 Person Search

Person search is a fast-evolving research topic. In 2014, [Xu et al., 2014] first introduced the person search problem and pointed out the conflicting nature between person detection and person identification sub-tasks. Person detection deals with common human appearance, while the identification task focuses on a person’s uniqueness. After [Xiao et al., 2017] introduced the first end-to-end person search framework in 2017, we have seen an increasing number of image-based person search works in the last three years. Meanwhile, in 2017, GNA-RNN [Li et al., 2017b] set the benchmark for text-based person search. We draw a timeline to present the person search works in Figure 3 and show the two divisions: image-based and text-based person search.

<table>
<thead>
<tr>
<th>Survey</th>
<th>Covering</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Islam, 2020]</td>
<td>Image-based</td>
<td>(Challenges: Solutions)</td>
</tr>
<tr>
<td>Ours</td>
<td>Image-based, Text-based</td>
<td>Discriminative person features: <strong>Deep feature representation learning</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Query-person gap:</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Deep metric learning</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Detection-identification inconsistency: <strong>Identity-driven detection</strong></td>
</tr>
</tbody>
</table>

Table 1: Summary of the main differences between the previous survey and ours. This survey focuses more on challenges and solutions.

Person search addresses person detection and person identification simultaneously. There are three significant person search challenges to be considered when developing a person search solution. Firstly, a person search model needs to learn discriminative person features from scene images suitable for matching the query identity. Inevitably, the learnt person features differ from the query identity features to some degrees. Therefore the second major challenge is how to bring the gap between the query and the detected person. The third challenge is related to the conflicting nature between person detection and person identification. Person detection deals with common person appearance, while the identification task focuses on a person’s uniqueness. The detected person may not be suitable for identity matching. For instance, a partial human body could be considered a person during detection and is inconsistent with the query identity at the identification stage, which may be a full person picture.

In this section, we analyse person search methods regarding above-mentioned three challenges and corresponding solutions from the following three aspects for both image-based and text-based person search:

- **Deep feature representation learning.** Addressing the challenge of learning discriminative person features from gallery images concerning background clutter, occlusion and poses etc.
- **Deep metric learning.** Addressing the challenge of bringing query-person gap by using loss functions to guide feature representation learning.
- **Identity-driven detection.** Addressing the challenge of mitigating the detection-identification inconsistency by incorporating query identities into the detection process.

#### 2.1 Deep Feature Representation Learning

Deep feature representation learning focuses on learning discriminative person features concerning distractors in the gallery images. The majority of the early methods exploited global person features, including context cues, while refining person proposals. Such as RCAA [Chang et al., 2018] utilises the relational spatial and temporal context in a deep reinforcement learning framework to adjust the bounding boxes constantly. However, these methods didn’t consider the background clutter in the proposal bounding boxes, resulting in a situation where different persons with similar backgrounds are close in the learnt feature space. SMG [Zheng...
et al., 2020a) eliminates background clutter using segmentation masks so that the learnt person features are invariant to the background clutter. NAE [Chen et al., 2020] separates persons and background by norms and discriminates person identities by angles. Person detection and object detection, in general, face the multi-scale matching challenge. To learn scale-invariant features, CLSA [Lan et al., 2018] and DHFF [Lu et al., 2019] utilise multi-level features from the identification network to solve the multi-scale matching problem with different multi-metric losses.

Local discriminative features are useful when two persons exhibit similar appearance and can’t be discriminated against merely by full-body appearance. APNet [Zhong et al., 2020] divides the body into six parts and uses an attention mechanism to weigh the body parts’ contribution further. Unlike APNet, which uses arbitrary body parts, CGPS [Yan et al., 2019] proposes a region-based feature learning model for learning contextual information from a person graph. BI-Net [Dong et al., 2020a] uses the guidance from the cropped person patches to eliminate the context influence outside the bounding boxes.

Deep feature representation learning in text-based person search learns visual representations for the detected person most correspondent to the textual features. Similar to image-based person search, text-based person search methods exploit global and local discriminative features. GNA-RNN [Li et al., 2017b] exploits global features in the first text-based LSTM-CNN person search framework and uses an attention mechanism to learn the most relevant parts. GNA-RNN only attends to visual elements and doesn’t address various text mechanism to learn the most relevant parts. GNA-RNN only attends to visual elements and doesn’t address various text

2.2 Deep Metric Learning

Deep metric learning tackles the query-person gap challenge with loss functions to guide the feature representation learning. The general purpose is to bring the detected person features close to the target identity while separating them from other identities. Similarity metrics such as Euclidean distance and cosine similarity are common measures to evaluate the similarity level among those query-person pairs. The identification task is generally formulated as a classification problem where conventional softmax loss trains the classifier. Softmax has a major problem of slow convergence with a large number of classes. OIM (Eq: 2) [Xiao et al., 2017] addresses this issue while exploiting large number of identities and unlabeled identities. OIAM [Gao et al., 2019] and IEL [Shi et al., 2018] further improve the OIM method with additional center losses. Different from OIM variances, I-Net [He and Zhang, 2019] introduces a Siamese structure with an online pairing loss (OPL) and hard example priority Softmax loss (HEP) to bring the query-person gap. RDLR [Han et al., 2019] uses the identification loss instead of regression loss for supervising the bounding boxes.

In the landmark OIM approach, the OIM loss effectively closes the query-person gap utilising labelled and unlabeled identities from training data. The probability of detected person features \( x \) being recognised as the identity with class-id \( i \) by a Softmax function:

\[
p_i = \frac{\exp(v_i^T x / \tau)}{\sum_{j=1}^{L} \exp(v_j^T x / \tau) + \sum_{k=1}^{Q} \exp(u_k^T x / \tau)}.
\]

Where \( v_i^T \) is the labelled person features for the \( i_{th} \) identity in the lookup table (LUT). \( v_j^T \) is the \( j_{th} \) labelled person features in the LUT. \( u_k^T \) is the \( k_{th} \) unlabelled person features in the LUT. \( \tau \) regulates probability distribution. OIM objective is to maximize the expected log-likelihood of the target \( t \):

\[
\mathcal{L} = \mathbb{E}_{x} \log p_{t}.
\]
Metric learning in text-based person search is to close the text-image modality gap. The main challenge in text-based person search is that it requires the model to deal with the complex syntax from the free-form textual description. To tackle this, methods like ViTAA, CMCE, PWM+ATH [Wang et al., 2020b; Li et al., 2017a; Chen et al., 2018b] employ attention mechanism to build relation modules between visual and textual representations. Unlike the above three methods, which are all the CNN-RNN frameworks, Dual Path [Zheng et al., 2020b] utilizes a cross-modal projection matching (CMPM) loss and a cross-modal projection classification (CMPC) loss to learn discriminative image-text representations. Similar to CMPM+CMPC, MAN [Jing et al., 2020b] proposes cross-modal objective functions for joint embedding learning to tackle the domain adaptive text-based person search.

Inspired by the recent success of knowledge distillation [Hinton et al., 2015], instead of directly training detection and identification sub-nets, the two modules can be learnt from the pre-trained detection and identification models [Munjal et al., 2019b]. DKD [Zhang et al., 2020] focuses on improving the performance of identification by introducing diverse knowledge distillation in learning the identification model. Specifically, a pre-trained external identification model is used to teach the internal identification model. A simplified knowledge distillation process is illustrated in Figure 4.

2.3 Identity-driven Detection

The detection-identification inconsistency challenge in image-based person search is tackled by incorporating identities into the detection process. This means during training, ground-truth person identities are used to guide person proposals, or at search time, the query identity information is utilised to refine the bounding boxes. Person search tackles person detection and person identification challenges in one framework. Existing person search methods can be divided into two-stage and end-to-end solutions from the architecture perspective. In two-stage person detection, the detection and identification models are trained separately for optimal performance of both detection and identification models [Zhang et al., 2020; Loesch et al., 2019]. However, due to the detection-identification inconsistent issue, the separately trained models may not yield the best person search result. To address the inconsistency problem between the two branches, TCTS [Wang et al., 2020a] and IGPN+PCB [Dong et al., 2020b] exploit query information at search time to filter out low probable proposals. End-to-end methods share visual features between detection and identification and significantly decrease runtime. However, joint learning contributes to sub-optimal detection performance [Wang et al., 2020a], which subsequently worsen the detection-identification inconsistency problem. To address the problem. NPSM [Liu et al., 2017a] and QEEPS [Munjal et al., 2019a] leverage query information to optimise person proposals in detection process. Differ from the query-guided methods, RDLR [Han et al., 2019] supervises bounding box generation using identification loss. Therefore, proposal bounding boxes are more reliable. In person search settings, the query identity is present in gallery images. Therefore, all methods mentioned above essentially incorporate identities into the detection process.

Text-based person search faces less detection-identification inconsistency challenge since the proposal person is identified by text-image matching without comparing bounding boxes. Therefore, text-based person search mainly focuses on learning visual and language features and improving the matching accuracy. The majority of current text-based person search methods are end-to-end frameworks that consist of a CNN backbone for extracting visual elements and a bi-LSTM for learning language representations. The two modules are jointly trained to build word-image relations from the learnt visual and language feature representations. CMCE [Li et al., 2017a] is the only two-stage framework in which the stage-one CNN-LSTM network learns cross-modal features, and in stage-two, the CNN-LSTM network refines the matching results using an attention mechanism.

3 Datasets and Evaluation

3.1 Datasets

CUHK-SYSU [Xiao et al., 2017] and PRW [Zheng et al., 2017] are the most commonly used datasets for image-based person search. CUHK-SYSU contains 18,184 images, 8,432 person identities, and 99,809 annotated bounding boxes. The training set contains 11,206 images and 5,532 query identities. The test set has 6,978 images and 2,900 query identities. PRW dataset has 11,816 frames and 43,110 person bounding boxes. 34,304 people have identities ranging from 1 to 932, and the rest are assigned identities of -2. The PRW training set has 5,704 images and 482 identities, and the test set has 6112 pictures and 450 identities. LSPS [Zhong et al., 2020] is a new image-based person search dataset, in which a total number of 51,836 pictures are collected. 60,433 bounding boxes and 4,067 identities are annotated. LSPS has a substantially larger number of incomplete query bounding boxes, making it a specialised dataset to evaluate methods exploiting
local discriminative features. CUHK-PEDES is currently the only dataset for text-based person search. It contains 40,206 images of 13,003 identities and 80,440 textual descriptions. Each picture has 2 textual descriptions. The dataset is divided into three parts, 11,003 training individuals with 34,054 images of 13,003 identities and 80,440 textual descriptions. Each picture has 2 textual descriptions.

### 3.2 Evaluation Metrics

Cumulative matching characteristics (CMC top-K) and mean averaged precision (mAP) are the primary evaluation metrics for person search. In CMC, the top-K predicted bounding boxes are ranked according to the intersection-over-union (IoU) overlap with the ground-truths equal to or greater than 0.5. The mAP is a popular evaluation metric in object detection, in which an averaged precision (AP) is calculated for each query person, and then the final mAP is calculated as an average of all APs.

### 3.3 Performance Analysis

In this section, we summarise and analyse the evaluation results considering the three significant challenges in person search discussed earlier. We aim to present the influencing factors that contribute to the person search performance. We don’t discuss CNN backbones as modern CNN backbones such as ResNet50 and VGG are similar in performance and are mostly interchangeable in different methods.

We summarise the evaluation results of image-based person search methods in Table 2. We annotate feature types and loss functions used for metric learning along with the methods. Image-based person search faces the steep detection-identification inconsistency challenge. Therefore, we divide image-based person search methods into identity-driven detection and non-identity-driven detection methods to analyse the identity-driven detection solution’s effectiveness.

Methods specifically addressing the detection and identification inconsistency challenge, such as IGPN, RDLR and TCTS, outperform methods addressing the detection and identification separately. Methods exploiting fine-grained
discriminative features without considering the detection-
identification inconsistency challenge don’t have a clear edge
over methods using global features. Our interpretation is that
the query identity presents in the gallery images. Therefore,
the detected person needs to be consistent with the query
identity for better query-person matching. For example, if
the detected person features are free from noises, the query
should be free of noises. Loss functions play critical roles in
guiding feature representation learning, such as using a cen-
ter loss on top of the OIM loss to bring the same identities
closer and separate different identities. Knowledge distilla-
tion is a notably effective strategy in training the detection
and identification models. KD-OIM, KD-QEEPS and DKD
beat the corresponding baseline methods without knowledge distillation.

The performance of the text-based person search meth-
ods on CUHK-PEDES is summarised in Table 4. We in-
clude feature types and loss functions along with the methods.
Text-based person search is essentially a text-image matching
problem, and fine-grained discriminative features play a crit-
ical role in cross-modal matching. Recent methods exploit-
ing fine-grained discriminative features with novel loss func-
tions outperform methods using global features and vanilla
Cross-Entropy loss. Specifically, ViTAA [Wang et al., 2020b]
exploring local discriminative features via attribute-feature
alignment achieves the best search results.

4 Discussion and Future Directions

In this survey, we review the recent person search advances
covering both image-based and text-based person search. It
remains an open question on addressing the three significant
person search challenges, namely the discriminative features,
the query-person gap and the detection-identification incon-
sistency. Next, we discuss a few future research directions.

Multi-modal person search. Exiting works focus on
search by either image or text. None of them attempted a
multi-modal search approach, in which query image and
query text complement each other. Multi-modal person
search is handy when a partial person image is available
such as a passport-sized image. At the same time, the free
text provides the rest of the body appearance. Specifically,
the CUHK-PEDES dataset can be extended with annotated
bounding boxes. Thus CUHK-PEDES has both annotated

<table>
<thead>
<tr>
<th>Method</th>
<th>Feature</th>
<th>Loss</th>
<th>top-1</th>
<th>top-5</th>
<th>top-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>GNA-RNN [Li et al., 2017b]</td>
<td>global</td>
<td>Cross entropy</td>
<td>19.05</td>
<td>53.63</td>
<td></td>
</tr>
<tr>
<td>CMCE [Li et al., 2017a]</td>
<td>global</td>
<td>CMCE loss</td>
<td>25.94</td>
<td>60.48</td>
<td></td>
</tr>
<tr>
<td>PWM+ATH [Chen et al., 2018b]</td>
<td>global</td>
<td>Cross entropy</td>
<td>27.14</td>
<td>49.45</td>
<td>61.02</td>
</tr>
<tr>
<td>Dual-Path [Zheng et al., 2020b]</td>
<td>global</td>
<td>Ranking loss, Instance loss</td>
<td>44.40</td>
<td>66.26</td>
<td>75.07</td>
</tr>
<tr>
<td>CMPM+CMPC [Zhang and Lu, 2018]</td>
<td>global</td>
<td>CMPM, CMPC</td>
<td>49.37</td>
<td>79.27</td>
<td></td>
</tr>
<tr>
<td>LPS+MCCL [Liu et al., 2019]</td>
<td>global</td>
<td>MCCL</td>
<td>50.58</td>
<td>79.06</td>
<td></td>
</tr>
<tr>
<td>A-GANet [Liu et al., 2019]</td>
<td>global &amp; pose</td>
<td>Binary Cross Entropy</td>
<td>53.14</td>
<td>74.03</td>
<td>81.95</td>
</tr>
<tr>
<td>PMA [Jing et al., 2020a]</td>
<td>global</td>
<td>Cross Entropy, GAN Loss</td>
<td>55.41</td>
<td>77.56</td>
<td>84.78</td>
</tr>
<tr>
<td>TIMAM [Sarafianos et al., 2019]</td>
<td>global</td>
<td>Alignment loss</td>
<td>55.97</td>
<td>75.84</td>
<td>83.52</td>
</tr>
</tbody>
</table>

Table 4: Performance of text-based person search methods on the CUHK-PEDES dataset.

bounding boxes and textual descriptions, making it a suitable
candidate dataset for multi-modal person search.

Attribute-based person search. It is a big challenge for a
machine to learn complex sentence syntax. Attribute-based
person search method AHM [Dong et al., 2019] outper-
forms the text-based method GNA-RNN [Li et al., 2017b]
evaluated on cropped person images with attribute annota-
tions. The state-of-the-art text-based person search method
ViTAA [Wang et al., 2020b] decomposes textual description
to attributes to learn fine-grained discriminative features. At-
tribute annotated datasets may ease this process and subse-
quently improve text-based person search performance.

Zero-shot person search. Text-based person search is es-
sentially a zero-shot learning problem, in which the query
person is unseen in training. [Dong et al., 2019] formul-
ates the attribute-based person search as a Zero-Shot Learn-
ing (ZSL) problem. In zero-shot learning, zero training image
is available at training time, and only semantic representa-
tions such as textual descriptions are available to infer unseen
classes. Text-based person search can leverage the knowledge
of zero-shot learning, such as using adversarially generated
person features to augment training data.

5 Conclusion

In this survey, we provide a systematic review of the recent
works on person search. For the first time, we surveyed pa-
pers on text-based person search which is less investigated
than image-based person search. We briefly discuss highly
regarded methods from the perspective of challenges and sol-
utions. We summarise and compare person search methods’
performance and provide insights that a person search method
needs to address the joint challenge of discriminative fea-
tures, query-person gap, and detection-identification incon-
sistency. We finally discuss some future research directions
which may be of interest to incumbent and new researchers
in the field.

Acknowledgments

This paper is partially supported by Australian Research
Council (ARC) Discovery Early Career Award (DECRA),
and National Natural Science Foundation of China under No.
61972315.
References

Chang et al., 2018 | Xiaojun Chang, Po-Yao Huang, Yi-Dong Shen, Xiaodan Liang, Yi Yang, and Alexander G. Hauptmann. RCAA: Relational Context-Aware Agents for Person Search. pages 84–100, 2018.


Han et al., 2019 | Chuchu Han, Jiacheng Ye, Yunshan Zhong, Xin Tan, Chi Zhang, Changxin Gao, and Nong Sang. Re-ID Driven Localization Refinement for Person Search. pages 9814–9823, 2019.


Li et al., 2019a | Jianheng Li, Fuhang Liang, Yuanxun Li, and Wei-Shi Zheng. Fast Person Search Pipeline. In 2019 IEEE International Conference on Multimedia and Expo (ICME), pages 1114–1119, July 2019. ISSN: 1945-788X.


Liu et al., 2019a | Jianheng Li, Fuhang Liang, Yuanxun Li, and Wei-Shi Zheng. Fast Person Search Pipeline. In 2019 IEEE International Conference on Multimedia and Expo (ICME), pages 1114–1119, July 2019. ISSN: 1945-788X.


Liu et al., 2018a | Hong Liu, Wei Shi, Weipeng Huang, and Qiao Guan. A Discriminatively Learned Feature Embedding Based on Multi-Loss Fusion For Person Search. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1668–1672, April 2018. ISSN: 2379-190X.


