RCAA: Relational Context-Aware Agents for Person Search

Xiaojun Chang[1], Po-Yao Huang[1], Yi-Dong Shen[2(✉)], Xiaodan Liang[1], Yi Yang[3], and Alexander G. Hauptmann[1]

[1] School of Computer Science, Carnegie Mellon University, Pittsburgh, USA
cxj273@gmail.com, berniebear@gmail.com, xdliang328@gmail.com, alex@cs.cmu.edu

[2] Institute of Software, Chinese Academy of Sciences, Beijing, China
ydshen@ios.ac.cn

[3] Centre for Artificial Intelligence, University of Technology Sydney, Ultimo, Australia
yee.i.yang@gmail.com

Abstract. We aim to search for a target person from a gallery of whole scene images for which the annotations of pedestrian bounding boxes are unavailable. Previous approaches to this problem have relied on a pedestrian proposal net, which may generate redundant proposals and increase the computational burden. In this paper, we address this problem by training relational context-aware agents which learn the actions to localize the target person from the gallery of whole scene images. We incorporate the relational spatial and temporal contexts into the framework. Specifically, we propose to use the target person as the query in the query-dependent relational network. The agent determines the best action to take at each time step by simultaneously considering the local visual information, the relational and temporal contexts, together with the target person. To validate the performance of our approach, we conduct extensive experiments on the large-scale Person Search benchmark dataset and achieve significant improvements over the compared approaches. It is also worth noting that the proposed model even performs better than traditional methods with perfect pedestrian detectors.

Keywords: Person search · Relational network

1 Introduction

Person re-identification (re-id) is an important research problem in which the goal is to match the same person across different camera views or across time within the same camera [1–5]. Its obvious applications include, but are not limited to, content-based video retrieval, video surveillance, and human-computer interaction [6]. Due to its importance for these applications, it has attracted increasing research attention in recent years. However, it remains challenging

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11213, pp. 86–102, 2018.
https://doi.org/10.1007/978-3-030-01240-3_6

and unsolved because of camera view changes, poor lighting conditions, severe background clutter and occlusion, and so on.

Despite the considerable progress that has been made, person re-id still cannot be directly applied to real-world applications. Most existing person re-id benchmark datasets and approaches focus on matching cropped person images from multiple non-overlapping cameras [7–9]. Although these approaches have achieved promising performance, they have major limitations for practical usage, since they are built upon the assumption of precise person detection. In real-world applications, precise bounding boxes are either unavailable or expensive to obtain. The off-the-shelf person detection algorithms would inevitably generate inaccurate proposals, thus deteriorating subsequent person re-id performance.

To close the gap between the research on person re-id and real-world applications, researchers have proposed the person search problem and several corresponding approaches [10–12]. We show the differences between person search and person re-id in Fig. 1. Xu *et al.* proposed to combine person detection and person matching scores using a sliding window search method [10]. Their method has several inherent drawbacks. Firstly, their algorithm is not scalable because of the sliding window algorithm. Secondly, they conduct person detection and search in two separate steps, which may lead to sub-optimal solution for person search. To address these problems, Xiao *et al.* proposed a new deep learning model to jointly conduct person detection and identification for person search [12]. However, their model also required to train a person proposal network for person candidate detection.

(a) Person Search (b) Person Re-ID

Fig. 1. We show examples of person search and person re-identification. Person search aims at finding a specified person from whole scene images, while person re-id aims to match cropped person images from multiple non-overlapping cameras. From the comparison, we can see that person search problem setting is closer to real-world applications and more challenging.

Spatial and temporal context may provide additional crucial information but still remains under-explored for person search. Spatial context has been proved useful in tasks like visual question answering [13]. The target-person-dependent spatial relationships between objects in the whole scene image may contribute to more discriminative representations. Additionally, the success of sequential decision making in object detection [14] also sheds light for person re-id. An agent making multi-step inferences may better locate the target person with consideration of its temporal action and state memory.

In this work, we propose a top-down search strategy powered by a spatial and temporal context-aware agent to address the limitations and opportunities discussed above. Specifically, given the whole scene, its local image features, and the query image, we leverage a target-person-dependent relational network to extract the spatial context between objects. Then our deep reinforcement learning agent selects the best action to narrow down the precise location of the target person at each time step based on the spatial context and its temporal action and state memory. The selected action is expected to keep the target person within the target box while cutting off as much background as possible. In this paper, we define 14 actions to perform the transitions of the target box. This step is repeated until the optimal result is obtained (when the agent selects the action "Terminate"). The whole framework comprises no person proposal computing and is end-to-end trainable.

To summarize, we make the following contributions to the field of person search.

- We make the earliest attempt to solve the person search problem as a conditional decision-making process and build the first deep reinforcement learning based person search framework.
- The proposed model is trained in an end-to-end fashion without proposal computing, which could be redundant and noisy. It is interesting to notice that our model even perform better than traditional methods with perfect pedestrian proposal detectors.
- We incorporate relational spatial and temporal contexts into the training procedure, which guides the model to generate more informative "experience".

2 Related Works

Person Re-identification. Pioneer researchers have proposed many algorithms to solve the re-id problem. These algorithms can be separated into two groups, discriminate feature learning [7,15–17] and distance metric learning [1,3,18,19]. The discriminate feature learning methods aim to learn distinct and informative features from cropped pedestrian images, while the distance metric learning methods usually learn distance metrics that are robust to sample variance.

Inspired by the phenomenal results achieved by deep learning networks in many computer vision applications [20–22], many researchers have explored different deep convolutional neural network (DCNN) models to solve the person re-id problem. Some researchers have employed a Siamese convolutional network

[23] for person re-id. For example, Ahmed *et al.* [24] and Li *et al.* [8] proposed using pairs of cropped pedestrian images as input and training the network using a binary verification loss function. Other researchers have adopted a triplet framework to improve person re-id performance. Ding *et al.* [25] and Cheng *et al.* [2] trained networks with triplet samples to make the features from the same pedestrian close and the features from different pedestrians far apart. We also notice that Zheng *et al.* also contributed a benchmark dataset for person search [5]. However, they proposed separate detection and re-id methods with scores re-weighting to solve the problem, while we propose a reinforcement learning framework for joint detection and re-id.

Pedestrian Detection. Early works on pedestrian detectors were built upon hand-crafted features and linear classifiers. Representative works include DPM [26], ACF [27] and Checkerboards [28]. These off-the-shelf pedestrian detectors are widely used for a variety of computer vision applications. Recently, various deep learning models have been proposed to boost the performance of pedestrian detection. For example, Cai *et al.* [29] proposed to seek an algorithm for optimal cascade learning under a criterion that penalizes both detection errors and complexity. Tian *et al.* [30] sought to jointly optimize pedestrian detection with semantic tasks, including pedestrian attributes and scene attributes. Ouyang *et al.* [31] proposed to handle occlusion by jointly learning features and the visibility of different body parts. They could effectively estimate the visibility of parts at multiple layers and learn their relationship with the proposed discriminative deep model. Luo *et al.* [32] propose to automatically learn hierarchical features, salience maps, and mixture representations of different body parts. Their model is able to explicitly model the complex mixture of visual variations at multiple levels.

Deep Reinforcement Learning. Deep reinforcement learning (DRL) has attracted much research attention over the last few years. Its goal is to learn a policy function that determines sequential actions by maximizing the cumulative future rewards [33]. Many researchers have attempted to incorporate deep neural networks with RL algorithms [34,35]. A common method is to use deep neural networks to represent RL models. These researchers have achieved human-level performance while playing Atari games [34] or Go [35]. Concurrently, some researchers propose to apply DRL to computer vision tasks, such as action recognition [36], object localization [14] and visual tracking [37].

Two widely used DRL methods are discussed in the literature, Deep Q-Networks (DQN) and policy gradient. As an exemplar of Q-learning, DQN approximates the state-action value function with deep neural networks. The network is trained by minimizing the temporal-difference errors [34]. To obtain better performance and maintain stability, researchers have proposed different network architectures based on DQN, *i.e.* Double DQN [38], DDQN [39], *etc.*

The goal of policy gradient methods is to use gradient descent to directly learn the policy by optimizing the deep policy networks with respect to the expected future reward. Williams *et al.* [40] proposed using the immediate reward to obtain an estimation of the policy value. They called this method REINFORCE and applied it to detect actions in videos.

3 Relational Context-Aware Agents

3.1 Overview

Person search solves the problem of finding the precise position of the target person from a gallery of whole scene images. The system dynamically locates the target person by sequential actions that are determined by a spatial-temporal context-aware agent. Our agent accepts the spatial and temporal context, the local image feature and the query image as input, and predicts the best action to take. The bounding box is transformed from its current state by the predicted action, and the next action is predicted from the next state. This process is repeated until we reach an optimal result. We show structure of the proposed model in Fig. 2.

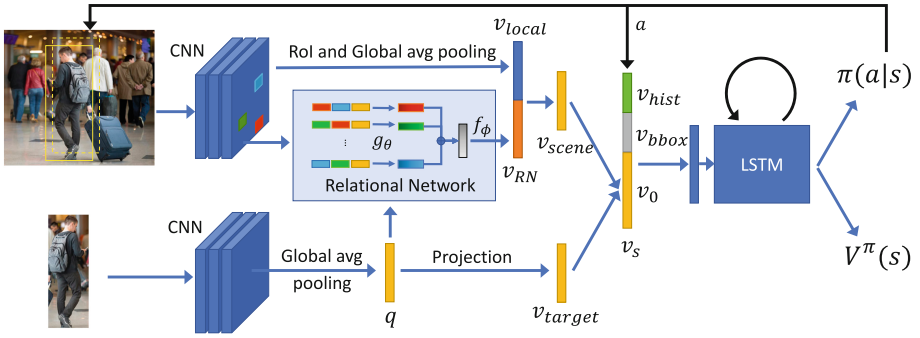


Fig. 2. The proposed relational spatial and temporal context-aware network. We adopt the relational network to compute the relational spatial context, and condition its processing on the target person. We encode the temporal context in the state of the LSTM. We use blue line to denote normal one-step feed-forward, and black line to denote action feedback loop.

3.2 Relational Decision Making

The proposed model follows Markov Decision Process (MDP), which is well suitable for modelling the sequential decision process. MDP is denoted as a tuple of states (S, A, R, γ), where $s \in S$ denotes a state of the environment, $a \in A$ denotes an action that the agent selects to transform the environment, $R : S \times A$ denotes the reward function that maps a state-action pair (s, a) to

a reward $r \in \mathbb{R}$, and $\gamma \in (0, 1]$ denotes a discount factor determining the decay rate in calculating the cumulative discounted reward of the entire trajectory. We represent the state and action as s_t and a_t, for $t = 1, \cdots, T$, where T denotes the termination step. We define the bounding box as $[x_t, y_t, w_t, h_t]$, where (x_t, y_t) is the center position, w_t and h_t are the width and the height of the bounding box, respectively.

Action. We show the 14 actions in Fig. 3. These actions can be grouped into three categories. The first category is for translating the current bounding box locally. The second category is for scaling the current bounding box to a smaller size (0.55 times as the original bounding box). The last category has only one action, namely "Terminate", which means the optimal result has been achieved. Similar to [14], the local translation group includes moving right/left horizontally, moving up/down vertically, making fatter/thinner horizontally and making fatter/thinner vertically. Each transformation action makes a discrete change to the current bounding box by a factor δ, where $\delta \in (0, 1]$. For example, if the action "moving right horizontally" is selected, the bounding box will change from $[x_t, y_t, w_t, h_t]$ to $[x_t + \delta * w_t, y_t, w_t, h_t]$. δ is set as 0.2 in our experiments, since this value has been selected in the literature for its good trade-off between speed and localization accuracy.

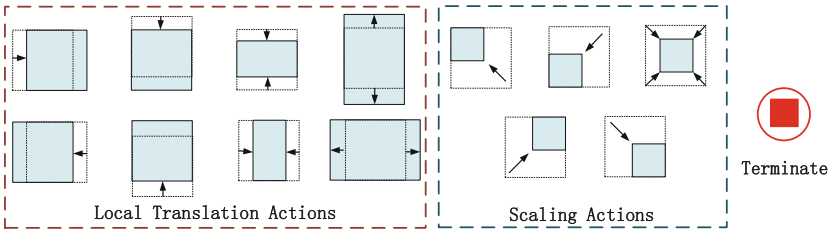


Fig. 3. Illustration of the 14 actions for the agent. The blue window with solid line denotes the bounding box after taking the corresponding action. The dashed line represents the bounding box before the action. The termination action means optimal result is reached

State. At each time step, we define the environment state as a combination of the whole image, the current bounding box, and the query person. In this paper, we initialize the bounding box as the whole image. We extract these features using the ResNet-50 [22] which is pre-trained on the large-scale ImageNet dataset [41]. Our framework utilizes the convolutional layers from the ResNet-50, followed by a ROI pooling layer [42,43] to extract a set of feature maps for the region proposal. We feed the whole image into the network, and obtain image feature maps after several convolutional layers. Then we feed these feature maps into a ROI pooling layer to extract the corresponding features for the object proposals.

Note that this is computationally efficient since we can share the computations in convolutional layers for all region proposals in one image. For the whole image and the query person image, we feed the feature maps into global average pooling layers [22] to obtain the feature vectors.

Reward. The reward function $R(s_t, a_t)$ is the improvement of localization performance when the action a_t is taken at the state s_t. In the literature [43], researchers employ the Intersection-over-Union (IoU) between the current bounding box and the ground-truth as the evaluation metric for its simplicity and effectiveness. However, as claimed in [44], this simple reward function tends to mislead the agent into learning suboptimal policies. It is essential to incorporate a shaping reward function into the original reward function. The shaping reward function is defined as:

$$F(s_t, a_t) = \gamma \Phi(s_{t+1}) - \Phi(s_t), \tag{1}$$

$$\Phi(s) = IoU(s_t) \tag{2}$$

where $\Phi(s)$ is a potential based reward. Following the literatures on deep reinforcement learning [43,45], we set the discount factor γ to 0.99. Ng *et al.* has proved that F is a necessary and sufficient condition to guarantee consistency with the optimal policy [44].

When the local translation action or scaling action is selected, by incorporating the shaping reward function, we have the following reward function:

$$R(s_t, a_t) = R'(s_t, a_t) + F(s_t, a_t), \tag{3}$$

where

$$R'(s_t, a_t) = \begin{cases} \text{IoU}(s_{t+1}), & \text{if IoU}(s_{t+1}) > \max \text{IoU}(s_k)|_{k=0}^{t} \\ -p, & \text{otherwise} \end{cases}. \tag{4}$$

The basic reward function $R'(s_t, a_t)$ will return $\text{IoU}(s_{t+1})$ when the new state s_{t+1} has higher IoU value than any previous states. Otherwise, we will give a penalty of $-p$ to the agent. Empirically, we set p as 0.05.

When the "Terminate" action is selected, the agent will receive a positive reward η if $\text{IoU}(s_T) > \tau$. Otherwise, a penalty will be given to the agent. The reward function $R(s_t, a_t)$ is defined as follows:

$$R(s_t, a_t) = \begin{cases} \eta & \text{if IoU}(s_t) > \tau \\ -\eta & \text{otherwise} \end{cases}. \tag{5}$$

In this paper, τ and η are empirically set as 0.5 and 1.0, respectively.

3.3 Network Structure

Following the traditional reinforcement learning setting, our agent interacts with the environment at each time step. We define a value function as $V^{\pi}(s) = \mathbb{E}[R_t|s_t = s]$, which measures the expected cumulative reward R_t for following a policy function π from any state s. The policy function $\pi(a|s)$ is used to select an action s from a set of actions given a state s. We approximate the value function and the policy function using a multi-layer neural network, which is a common practice in DRL. The network has two outputs, *i.e.* the distribution $\pi(a|s)$ over the possible actions and value estimation $V^{\pi}(s)$.

We employ the ResNet-50 [22] for feature extraction because it has demonstrated its superiority in terms of person re-id [46]. To encode the relational context information, we incorporate a Relational Network (RN) [13] to consider all the relations across all pairs of objects in the image. The RN can be simply defined as follows:

$$\mathrm{v_{RN}} = f_{\phi}(\sum_{i,j} g_{\theta}(o_i, o_j)), \tag{6}$$

where $O = (o_1, \cdots, o_n)$ denote a set of "objects", o_i is the i-th object, and f_{ϕ} and g_{θ} are functions with parameters ϕ and θ. Santoro *et al.* have demonstrated that CNN embeddings can be used as a set of objects for an RN. We feed the image into the ResNet-50 and get the k feature maps of size $d \times d$ from the final convolutional layer, where k is the number of kernels in the final convolutional layer. We tag each of the d^2 k-dimensional cells in the $d \times d$ feature maps with an arbitrary coordinate indicating its relative spatial position, and treat it as an object for the RN. The existence and meaning of an object-object relation should be relevant to the query person. Hence, we make the function g_{θ} condition its processing on the query person:

$$v_{RN} = f_{\phi}(\sum_{i,j} g_{\theta}(o_i, o_j, q)). \tag{7}$$

We concatenate the whole scene image representation v_{RN} and the local representation v_{local}, resulting in the scene representation v_{scene}. We feed the feature maps of the target person to global average pooling and achieve the target person representation v_{target}. We project both features into a ℓ_2-normalized 256 dimensional subspace, and apply dot product and ℓ_2-norm to v'_{scene} and v'_{target}. For simplicity, we drop the apostrophe. Hence, we encode the observation of the current state as:

$$v_0 = \frac{v_{\mathrm{scene}} \cdot v_{\mathrm{target}}}{||v_{\mathrm{scene}} \cdot v_{\mathrm{target}}||}, \tag{8}$$

where $\cdot$ denotes element-wise dot product.

To step further, the temporal context is explored to track the states that the agent has encountered as well as all the actions that the agent has taken. In this work, we record 50 previous actions, resulting in a history vector $v_{\mathrm{history}} \in \mathbb{R}^{700}$.

We encode the relative location and size of the region using a 5-dimensional vector:

$$v_{\text{bbox}} = \left[\frac{x_t - \frac{w_t}{2}}{w_t}, \frac{y_t - \frac{h_t}{2}}{h_t}, \frac{x_t + \frac{w_t}{2}}{w_t}, \frac{y_t + \frac{h_t}{2}}{h_t}, \frac{S_{\text{bbox}}}{S_{\text{image}}} \right], \tag{9}$$

where S_{bbox} is the size of bounding box and S_{image} is the size of the image. Then we represent the state as a vector $v_s = [v_0, v_{\text{history}}, v_{\text{bbox}}]$. We pass the vector representation of the state v_s to two FC layers with the same output size of 1,024, following by using a Long Short-Term Memory (LSTM) cell with Layer Normalization to track the past states. The temporal context for subsequent decision making is encoded by the state inside the LSTM cell.

3.4 Training

Although using a single agent to collect experiences may obtain promising results, it may achieve highly correlated data. Updating the system with these data would lead the agent to learn a suboptimal solution. To avoid the suboptimal solution, we employ the asynchronous advantage actor-critic (A3C) method [45] which asynchronously execute multiple agents in parallel, on multiple instances of the environment. A3C contains a policy function $\pi(a|s;\theta_\pi)$ and an estimate of the value function $V(s;\theta_v)$, where θ_π and θ_v are the parameters of the policy function and the value function, respectively. When we process one query, an agent interacts with the environment constructed by the query using the current network, and generates an episode $\{(s_t, a_t, r_t)\}_{t=0,\cdots,T}$ for training. The query is selected randomly during training. We update the network parameters asynchronously.

Following [45], we empirically group every N consecutive experiences in every episode. At each time step t, we convert the reward as

$$R'_t = \sum_{k=t}^{t_m(t)-1} \gamma^{k-t} r_k + \gamma^{t_m(t)-t} V(s_{t_m}(t)), \tag{10}$$

if the condition $t + N \leq T$ is met. Note that $t_m(t) = \lceil \frac{t}{N} \rceil \cdot N$. Otherwise, we convert the reward according to $R'_t = \sum_{k=t}^{T} \gamma^{k-t} r_k$. We collect all the tuples in parallel, and use them to optimize in batch mode. We train the network by the ADAM optimizer [47], and optimize in batch mode as follows:

$$\begin{aligned} \theta_\pi \leftarrow \theta_\pi + \alpha((R'_t - V(s_t;\theta_v)) \bigtriangledown_{\theta_\pi} \log \pi(a_t|s_t;\theta_\pi) \\ + \beta \bigtriangledown_{\theta_\pi} H(\pi(\cdot|s_t;\theta_\pi))) \end{aligned} \tag{11}$$

$$\theta_v \leftarrow \theta_v - \alpha \bigtriangledown_{\pi_v} (R_t - V(s_t;\theta_v))^2, \tag{12}$$

where α denotes the learning rate, $H(\pi(\cdot|s_t;\theta_\pi))$ represents the entropy of the policy [45], β is the hyper parameter, $(R'_t - V(s_t;\theta_v)) \bigtriangledown_{\theta_\pi} \log \pi(a_t|s_t;\theta_\pi)$ is

policy gradient which calculates the direction to update the policy such that the rewards of the agent will be improved.

4 Experiments

To evaluate the performance of the proposed model and study the impact of various factors on person search performance, we conduct extensive experiments on the large-scale person search dataset. In this section, we first describe the detailed experimental setup in Sect. 4.1. Then we compare the proposed model with the baseline algorithms in terms of Cumulative Matching Characteristics (CMC Top-K) and mean average precision (mAP). Afterwards, we conduct ablation study to analyze the effects of different components. Finally, we study the influence of gallery size.

4.1 Experimental Setup

Implementation Details: We use PyTorch to implement our model, and run the experiments on the NVIDIA TITAN Xp GPU. During training, 50 separate processes are used to run agents with environments, and one process to run policy and value network. When the training is finished, we fix the policy and value network for testing. A single agent is used to process each query for testing. For each query, we rank all the value V and retrieve the top ranked results.

Dataset Description: We test on the large-scale person search benchmark dataset provided by [12]. To the best of our knowledge, this is the only dataset available for person search. This dataset contains 8,432 labeled identities, who appear across different images. These people appear with full bodies and normal poses. Since person search problem mainly rely on body shapes and clothes rather than faces, the authors did not annotate people who change clothes and decorations in different video frames. This dataset has rich variations of pedestrian scales. The dataset is officially split into a training and a testing subset, without overlapping images between them. The test identity instances are divided into queries and galleries. We show the statistics of this dataset in Table 1.

Table 1. Statistics of the person search dataset with respect to training/test splits.

Split	# Images	# Pedestrians	# Identities
Training	11,206	55,272	5,532
Testing	6,978	40,871	2,900
Overall	18,184	96,143	8,432

Evaluation Protocols and Metrics: Following [12], we use two evaluation metrics to measure the performance, namely cumulative matching characteristics (CMC top-K) and mean averaged precision (mAP). The first metric has been widely used for the person re-id problem, where a matching is counted if there is at least one of the top-K predicted bounding boxes overlaps with the ground truths with intersection-over-union (IoU) greater or equal to 0.5. The second metric has been commonly used in the object detection tasks. The ILSVRC object detection criterion is used to judge the correctness of predicted bounding boxes. We calculate an averaged precision (AP) for each query based on the precision-recall curve, and then average the APs across all the queries to get the final result.

Compared Algorithms: To demonstrate the performance of our model, we first compare with the conventional methods for person search. These methods assume perfect pedestrian detection and break person search down into two separate tasks. We use two pedestrian detection methods and five pedestrian re-id approaches in the experiments. We use the off-the-shelf deep learning CCF detector [48] and Faster-RCNN (CNN) [49] with ResNet-50, specifically fine-tuned on the person search dataset. The ground truth (GT) bounding boxes are also used as the results a perfect detector. For pedestrian re-id, we used several well-known feature representations in the field, namely DenseSIFT-ColorHist (DSIFT) [16], Bag of Words (BoW) [50] and Local Maximal Occurrence (LOMO) [51]. We use each of the feature representation together with a specific distance metric learning algorithm, namely Euclidean, Cosine similarity, KISSME [52] and XQDA [51].

We also compare with a joint detection and identification feature learning algorithm [12], which jointly handles pedestrian detection and person re-id in a single CNN. To the best of our knowledge, this is the state-of-the-art person search algorithm in the literature. We drop the pedestrian proposal network, and train the remaining net to classify identities with Softmax loss from cropped pedestrian images, resulting another baseline method (IDNet), which has been exploited in [53].

4.2 Performance Comparison

We report the experimental results in Tables 2 and 3. We first compare the proposed framework with the conventional person search algorithms that break down the problem into two steps. From the experimental results shown in Tables 2 and 3, we can observe that our model performs much better than the compared baseline algorithms. The experimental results indicate that the pedestrian detector has a great impact on each person re-id algorithm. For example, for DSIFT + Euclidean, if an off-the-shelf detector (CCF) is used instead of a perfect detector (GT), the performance drops from 45.9% to 11.7%. This phenomenon confirms that it does not make sense to directly apply off-the-shelf pedestrian

detector for the real-world person search problems. The incorrect detection result of the detector will deteriorate the subsequent re-id performance.

Table 2. Experimental comparisons for person search on the large-scale benchmark dataset. Cumulative matching characteristics (CMC top-K) is used as the evaluation metric. Results are shown in percentages. Larger CMC indicates better performance. The best results are marked in bold.

	CCF				CNN				GT			
	Top-1	Top-5	Top-10	Top-20	Top-1	Top-5	Top-10	Top-20	Top-1	Top-5	Top-10	Top-20
DSIFT + Euclidean	11.7	31.4	45.8	63.9	39.4	65.2	77.6	81.8	45.9	67.2	78.1	86.3
DSIFT + KISSME	13.9	34.2	48.7	66.4	53.6	68.8	78.5	86.4	61.9	74.2	83.5	88.7
BoW + Cosine	29.3	54.2	71.5	86.8	62.3	74.7	82.1	88.2	67.2	76.8	85.8	89.9
LOMO + XQDA	46.4	67.2	78.5	87.6	74.1	79.8	85.6	91.1	76.7	84.7	88.4	92.2
IDNet	57.1	80.2	90.1	95.6	74.8	80.7	87.9	93.0	78.3	85.6	89.1	94.3
Joint Detec. & Identifi.	–	–	–	–	78.7	83.6	90.5	95.2	80.5	87.8	91.2	96.3
RCAA	–	–	–	–	**81.3**	**88.2**	**92.4**	**97.6**	–	–	–	–

Table 3. Experimental comparisons for person search on the large-scale benchmark dataset. Mean average precision (mAP) is used as the evaluation metric. Results are shown in percentages. Larger mAP indicates better performance. The best results are marked in bold.

mAP (%)	CCF	CNN	GT
DSIFT + Euclidean	11.3	34.5	41.1
DSIFT + KISSME	13.4	47.8	56.2
BoW + Cosine	26.9	56.9	62.5
LOMO + XQDA	41.2	68.9	72.4
IDNet	50.9	68.6	73.1
Joint Detec. & Identifi.	–	75.7	77.9
RCAA	–	**79.3**	–

We can also find that the proposed model outperforms the joint detection and identification method by a large margin. For example, the proposed model outperforms Joint Detec. & Identifi. with mAP of 79.3 *vs* 75.7 on the benchmark dataset. We attribute this improvement to the end-to-end model and its ability to exploit relational context in the visual data. Also, since the proposed model is proposal-free, it is more efficient than the proposal-based methods.

Interestingly, we notice that the proposed model even outperforms the baseline algorithms using perfect pedestrian detectors (GT), which further confirms the superiority of our model for person search problem. For example, the proposed model outperforms Joint Detect. & Identifi. (with ground truth) with mAP of 79.3 *vs* 77.9.

We show some examples of our result on the benchmark dataset in Fig. 4. From the examples we can see that given a target person, the system can correctly retrieve and localize the required person from the gallery set.

Fig. 4. Example from the testing set of the benchmark dataset. The ground truth is the yellow bounding box. And the predicted box is the blue box. Best viewed in color.

4.3 Ablation Study

In this section, we conduct experiments to test the effect of the context, *i.e.* the relational spatial context and the temporal context, in the reinforcement learning algorithm. Note that the proposed model explores the relational context using a Relational Network, and the temporal context using an LSTM. We train two modified versions of the proposed model. The first one does not use the spatial and temporal context, which is denoted as "w/o spatial & temporal". The second does not consider the spatial context, which is denoted as "w/o spatial". The experimental results are reported in Table 4. From the experimental results we observe that both the spatial and temporal context plays a vital role in the proposed model.

We also compare with the global context proposed in [22]. We feed the feature maps into a RoI pooling layer, following by a global average pooling layer, resulting in global feature. We replaced the relational feature with the global feature, which is denoted as "w. global + temporal". The experimental results shown in Table 4 confirms that relational context achieves better performance than global context for person search problem.

4.4 The Influence of Gallery Size

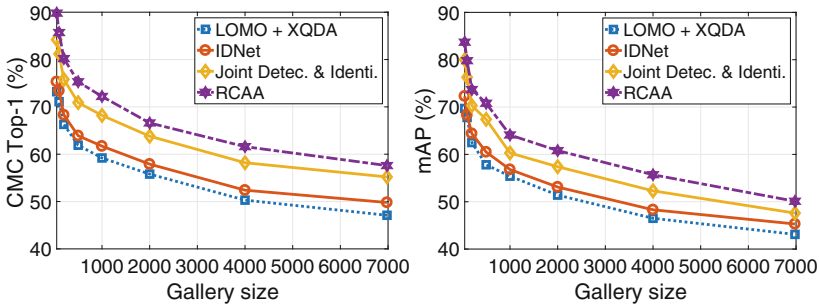
Intuitively, the person search problem will become extremely challenging when the gallery size increases sharply. In this section, we vary the gallery size from 50 to full set of 6,978 images to test the influence of gallery size. We report the experimental results in terms of CMC top-1 and mAP in Fig. 5. When we

Table 4. Comparison between our model and other variants for person search problem. CMC Top-K and mAP are reported in this table. Performance is reported in percentages. Larger number indicates better performance. The best performance is marked in bold.

	CMC Top-K				mAP
	Top-1	Top-5	top-10	Top-20	
(w/o both)	72.5	79.4	83.2	86.9	69.8
(w/o relational spatial)	74.8	81.6	85.5	88.2	71.4
(w. global + temporal)	78.9	85.7	89.4	93.6	76.7
(Ours full)	**81.3**	**88.2**	**92.4**	**97.6**	**79.3**

process each query, we randomly select the corresponding gallery images from the full set.

From the experimental results, we have the following observations: (1) as the gallery size increases, the performance of all the compared algorithms decreases; (2) the proposed model outperforms the other compared algorithms by a large margin with right to different gallery sizes.



(a) CMC Top-1 (b) mAP

Fig. 5. The performance variance with right to different gallery sizes. The performance is reported in percentages.

5 Conclusions and Future Work

In this paper, we have made the earliest attempt to address the person search problem and built the first deep reinforcement learning based person search framework. Unlike previous works which rely on pedestrian proposal net, our approach leverages the relational context information and exploits the visual information and the query person a priori in a joint framework. We have conducted extensive experiments to evaluate the performance of our model. The experimental results confirm its superiority.

In the future we plan to exploit lenient learning [54] in our framework as stored transitions can become outdated due to agents updating their respective policies in parallel.

Acknowledgements. This work was supported in part by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number D17PC00340, in part by China National 973 program 2014CB340301, and in part by the Data to Decisions CRC (D2D CRC) and the Cooperative Research Centres Programme. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation/herein. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

References

1. Cheng, D., Chang, X., Liu, L., Hauptmann, A.G., Gong, Y., Zheng, N.: Discriminative dictionary learning with ranking metric embedded for person re-identification. In: IJCAI (2017)
2. Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In: CVPR (2016)
3. Zheng, W., Gong, S., Xiang, T.: Person re-identification by probabilistic relative distance comparison. In: CVPR (2011)
4. Zheng, L., et al.: MARS: a video benchmark for large-scale person re-identification. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 868–884. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_52
5. Zheng, L., Zhang, H., Sun, S., Chandraker, M., Tian, Q.: Person re-identification in the wild. CoRR abs/1604.02531 (2016)
6. Yu, S., Yang, Y., Hauptmann, A.G.: Harry potter's marauder's map: localizing and tracking multiple persons-of-interest by nonnegative discretization. In: CVPR (2013)
7. Li, W., Wang, X.: Locally aligned feature transforms across views. In: CVPR (2013)
8. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: deep filter pairing neural network for person re-identification. In: CVPR (2014)
9. Zheng, W., Gong, S., Xiang, T.: Associating groups of people. In: BMVC (2009)
10. Xu, Y., Ma, B., Huang, R., Lin, L.: Person search in a scene by jointly modeling people commonness and person uniqueness. In: MM. ACM (2014)
11. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: End-to-end deep learning for person search. CoRR abs/1604.01850 (2016)
12. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: CVPR (2017)
13. Santoro, A., et al.: A simple neural network module for relational reasoning. CoRR abs/1706.01427 (2017)
14. Caicedo, J.C., Lazebnik, S.: Active object localization with deep reinforcement learning. In: ICCV (2015)

15. Wang, X., Doretto, G., Sebastian, T., Rittscher, J., Tu, P.H.: Shape and appearance context modeling. In: ICCV (2007)
16. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: CVPR (2013)
17. Khamis, S., Kuo, C.-H., Singh, V.K., Shet, V.D., Davis, L.S.: Joint learning for attribute-consistent person re-identification. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014. LNCS, vol. 8927, pp. 134–146. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16199-0_10
18. Liao, S., Li, S.Z.: Efficient PSD constrained asymmetric metric learning for person re-identification. In: ICCV (2015)
19. Pedagadi, S., Orwell, J., Velastin, S.A., Boghossian, B.A.: Local fisher discriminant analysis for pedestrian re-identification. In: CVPR (2013)
20. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
23. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Deep metric learning for person re-identification. In: ICPR (2014)
24. Ahmed, E., Jones, M.J., Marks, T.K.: An improved deep learning architecture for person re-identification. In: CVPR (2015)
25. Ding, S., Lin, L., Wang, G., Chao, H.: Deep feature learning with relative distance comparison for person re-identification. Pattern Recogn. **48**(10), 2993–3003 (2015)
26. Felzenszwalb, P.F., Girshick, R.B., McAllester, D.A., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
27. Dollár, P., Appel, R., Belongie, S.J., Perona, P.: Fast feature pyramids for object detection. IEEE Trans. Pattern Anal. Mach. Intell. **36**(8), 1532–1545 (2014)
28. Zhang, S., Benenson, R., Schiele, B.: Filtered channel features for pedestrian detection. In: CVPR (2015)
29. Cai, Z., Saberian, M.J., Vasconcelos, N.: Learning complexity-aware cascades for deep pedestrian detection. In: ICCV (2015)
30. Tian, Y., Luo, P., Wang, X., Tang, X.: Pedestrian detection aided by deep learning semantic tasks. In: CVPR (2015)
31. Ouyang, W., Wang, X.: A discriminative deep model for pedestrian detection with occlusion handling. In: CVPR
32. Luo, P., Tian, Y., Wang, X., Tang, X.: Switchable deep network for pedestrian detection. In: CVPR (2014)
33. Sutton, R.S.: Introduction to Reinforcement Learning, vol. 135
34. Mnih, V., et al.: Playing atari with deep reinforcement learning. CoRR (2013)
35. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., et al.: Mastering the game of go with deep neural networks and tree search. Nature **529**(7587), 484–489 (2016)
36. Jayaraman, D., Grauman, K.: Look-Ahead before you leap: end-to-end active recognition by forecasting the effect of motion. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 489–505. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_30
37. Yun, S., Choi, J., Yoo, Y., Yun, K., Choi, J.Y.: Action-decision networks for visual tracking with deep reinforcement learning. In: ICCV (2017)

38. van Hasselt, H., Guez, A., Silver, D.: Deep reinforcement learning with double Q-learning. In: AAAI (2016)
39. Wang, Z., Schaul, T., Hessel, M., van Hasselt, H., Lanctot, M., de Freitas, N.: Dueling network architectures for deep reinforcement learning. In: ICML (2016)
40. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach. Learn. **8**, 229–256 (1992)
41. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
42. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1904–1916 (2015)
43. Jie, Z., Liang, X., Feng, J., Jin, X., Lu, W., Yan, S.: Tree-structured reinforcement learning for sequential object localization. In: NIPS (2016)
44. Ng, A.Y., Harada, D., Russell, S.J.: Policy invariance under reward transformations: theory and application to reward shaping. In: Proceedings of the Sixteenth International Conference on Machine Learning (ICML 1999), Bled, Slovenia, 27–30 June 1999, pp. 278–287 (1999)
45. Mnih, V., et al.: Asynchronous methods for deep reinforcement learning. In: ICML (2016)
46. Geng, M., Wang, Y., Xiang, T., Tian, Y.: Deep transfer learning for person re-identification. CoRR abs/1611.05244 (2016)
47. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: ICLR (2015)
48. Yang, B., Yan, J., Lei, Z., Li, S.Z.: Convolutional channel features. In: ICCV (2015)
49. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: NIPS (2015)
50. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: ICCV (2015)
51. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: CVPR (2015)
52. Köstinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: CVPR (2012)
53. Wang, Z., Li, H., Ouyang, W., Wang, X.: Learning deep representations for scene labeling with semantic context guided supervision. CoRR abs/1706.02493 (2017)
54. Potter, M.A., De Jong, K.A.: A cooperative coevolutionary approach to function optimization. In: Davidor, Y., Schwefel, H.-P., Männer, R. (eds.) PPSN 1994. LNCS, vol. 866, pp. 249–257. Springer, Heidelberg (1994). https://doi.org/10.1007/3-540-58484-6_269