Person Search in a Scene by Jointly Modeling People Commonness and Person Uniqueness

Yuanlu Xu
Sun Yat-Sen University
eng.yuanlu.xu@ieee.org

Bingpeng Ma
Univ. of China Academy Sci.
bpma@ucas.ac.cn

Rui Huang[*]
Huazhong Univ. of Sci.&Tech.
ruihuang@hust.edu.cn

Liang Lin
Sun Yat-Sen University
linliang@ieee.org

ABSTRACT

This paper presents a novel framework for a multimedia search task: searching a person in a scene using human body appearance. Existing works mostly focus on two independent problems related to this task, i.e., people detection and person re-identification. However, a sequential combination of these two components does not solve the person search problem seamlessly for two reasons: 1) the errors in people detection are carried into person re-identification unavoidably; 2) the setting of person re-identification is different from that of person search which is essentially a verification problem. To bridge this gap, we propose a unified framework which jointly models the commonness of people (for detection) and the uniqueness of a person (for identification). We demonstrate superior performance of our approach on public benchmarks compared with the sequential combination of the state-of-the-art detection and identification algorithms.

Categories and Subject Descriptors

I.4 [**Image Processing and Computer Vision**]: Miscellaneous

General Terms

Theory; Algorithm; Experimentation

Keywords

person search; generative model; GMM; Fisher vector

1. INTRODUCTION

Over the past few decades, mature and low-cost vision acquisition equipments substantially boosted academic studies and industrial applications in vision related multimedia. One of the fastest growing areas is intelligent surveillance. From the perspective of public security, searching suspects and missing people in surveillance footage is an effective procedure to help solve cases. Hereby, we investigate this particular multimedia search problem, i.e., **person search**, aiming at locating a specific person in a scene given a query image using visual clues.

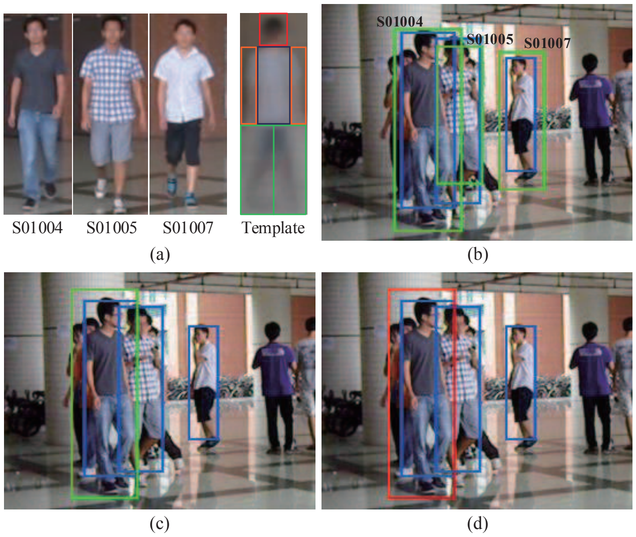

Figure 1: An illustration of searching a person in a scene. (a) Query images and template. (b) The results of our search approach. Blue rectangles are the groundtruth and green ones the correctly identified persons. (c) The people detection results from HOG-based detector. (d) The person re-identification fails to identify the correct person based on the detection results.

The most-researched clue for this task is face. Both face detection and recognition have achieved impressive performance through decades of research. However, due to the low resolution and pose variation of individuals in typical surveillance footage, faces are often deemed of limited usefulness. Going beyond faces, human body appearance is another important clue for searching a person in a scene. Nevertheless, there is a gap between the current research trends and the actual task we are facing.

Person search can be viewed as a special case of image search, or content-based image retrieval [10, 7], which has long been studied in the multimedia field. Traditional image search does not make use of the prior knowledge about

[*]Corresponding author. This work was supported in part by NSFC 61105014, 61332016, Hi-Tech R&D Program of China (no.2013AA013801), Special Project on Integration of Industry, Education and Research of Guangdong (no.2012B091100148), and Fundamental Research Funds for the Central Universities (no.13lgjc26).

MM'14, November 3–7, 2014, Orlando, Florida, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3063-3/14/11 ...$15.00.
http://dx.doi.org/10.1145/2647868.2654965.

the object category that the queried object belongs to, while in person search we know that the query image contains a person. The benefit is that we can now use this knowledge to facilitate our searching procedure, e.g., by incorporating the common characteristics of human body obtained from a training set besides the query image itself. Another difference is that person search is instance-level search that calls for a higher degree of accuracy for practical use than general image search.

Besides image search, extensive studies have been conducted on two other problems related to person search, i.e., people detection and person re-identification. People detection tries to find all the people in a scene, which is different from our mission of seeking a specific person given a query image. One of the most influential people detection methods employs the Histograms of Oriented Gradients (HOG) features and a linear SVM classifier [1]. The state-of-the-art general object detection methods such as Deformable Part Models (DPM) [4] can also be used for people detection. We refer the readers to a detailed review [2]. From the object detection point of view, person search is an extreme case of fine-grained object detection, i.e., instance-level detection.

To achieve the goal of person search, a further identification step is required to follow people detection, that is, a mode to match images of people using the whole body appearance, which is usually called person re-identification. Previous works on person re-identification can be roughly divided into the following two categories. *Learning-based* methods tend to learn the discriminative features [6], distance metric [14] or transformations [8] across views and cameras of the same person. Despite acknowledged success, learning-based methods suffer from model over-fitting and currently are unsuitable for large-scale data. *Matching-based* approaches, on the other hand, aim at localizing and comparing the corresponding image patches [9, 13], image segments [3] or body parts [12] in images.

However, there are two drawbacks in solving the person search task by sequentially combining people detection and person re-identification. First, person re-identification focuses on measuring the similarity of two windows (manually cropped or obtained by detection) containing the roughly aligned human subjects. Thus, the overall performance of identifying a specific person in a scene heavily relies on the performance of the people detector. Although good progress has been made on people detection, the results are not as promising when there are occluded people, as illustrated in Fig. 1. Obviously, a false positive or false negative reported by the people detector will unavoidably generate a corresponding false positive or false negative in the subsequent person re-identification results.

Second, the setting of person re-identification is different from the setting of person search. In current person re-identification research, for each window to be identified, a gallery set is commonly given so that the identity of the window can be classified or ranked according to the gallery. While in the person search task, for each candidate window, there is only one query image to be verified against. Therefore, many current person re-identification methods [6, 14] do not work in this setting.

These observations inspire us to treat the person search problem in a unified way rather than solving it in a sequential fashion. We hereby propose a framework to jointly optimize the detection of people and the identification of the queried person. The philosophy behind our proposal is that people detection algorithms use the most common features (**commonness**) shared by all human bodies to distinguish them from other objects, while the person re-identification algorithms need the most distinctive features (**uniqueness**) that are unique to a certain person to discriminate him/her against other people. Recent studies have shown that even small salient regions in appearance play an important role in identification [13]. However, in the sequential framework, the salient regions may easily be ignored by the people detector, especially if they belong to an occluded person, which results in false negatives in detection and consequently in the final results of search. On the other hand, if we only focus on the uniqueness of the queried person, we may be able to pick up the salient regions, but will probably pick up many false positives from the background as well.

In particular, our algorithm finds the queried person by testing a binary classifier in sliding windows throughout the image spatially and across scale levels, followed by non-maximum suppression to remove multiple responses of the classifier on the same individual at slightly shifted spatial locations and neighboring scales. This binary classifier is trained to verify whether the queried person is present in the current window or not and consists of two components: 1) a generative model representing the commonness of people using the average probabilistic distribution of the training data; 2) a Fisher vector feature [11] representing the uniqueness of a person, using the learned generative model as the kernel. The final verification is based on fusion of the commonness of the human body (to reduce false positives) and the uniqueness of the queried person (to reduce false negatives).

The main contributions are: 1) we argue that the simple combination of current solutions to the related problems (e.g., people detection and person re-identification) does not straightforwardly solve the new person search problem; 2) we propose a unified framework for person search by jointly modeling people commonness and person uniqueness, and show its superiority with experiments.

2. PROPOSED FRAMEWORK

In this section, we introduce a unified framework for person search. Given a query image I_q and a scene I_s, we want to find the possible location L the queried person stayed at. Since we know we are looking for a person, which belongs to the category of people, we can also make use of an additional set of images of people, T. The problem can be solved by maximizing a posterior

$$\begin{aligned} P(L|I_q, I_s, T) &\propto P(L|I_s, T)\, P(L|I_q, I_s) \\ &\propto e^{-D_{com}(L, I_s, T)} \cdot e^{-D_{uni}(L, I_q, I_s)}, \end{aligned} \tag{1}$$

because T is independent to I_q, and assuming a uniform prior for L. $D_{com}(\cdot)$ and $D_{uni}(\cdot)$ are two distance functions measuring respective commonness (Sec. 2.2) and uniqueness (Sec. 2.3), based on appearance computed locally (Sec. 2.1).

2.1 Local Descriptor

To take advantage of the common spatial configuration of human bodies (e.g., mostly standing upright, often appearing symmetric) without using sophisticated part matching algorithms, we describe the entire window with six smaller rectangles roughly corresponding to the six human body parts (i.e., head, torso, left and right arms, left and right legs), as shown in Fig. 1(a). The template is empirically derived from the average image of the training set. Feature extraction and model training are performed in each part separately. For clarity, we limit the following discussion in one part. The complete feature or model is a concatenation of the part features or models.

In order to capture the structure, color, gradient and orientation information, we use a very concise 7-dimensional low-level descriptor for each pixel [9]:

$$f(x,y) = [\tilde{x},\, \tilde{y},\, I(x,y),\, \frac{\partial I}{\partial x},\, \frac{\partial I}{\partial y},\, \frac{\partial^2 I}{\partial x^2},\, \frac{\partial^2 I}{\partial y^2}], \tag{2}$$

where $\tilde{x}$ and $\tilde{y}$ are the pixel relative coordinates within the part, $I(x,y)$ is the pixel intensity at position (x,y), $\frac{\partial I}{\partial x}$ and $\frac{\partial I}{\partial y}$ are the first-order derivatives of image I with respect to the horizontal and vertical directions, while $\frac{\partial^2 I}{\partial x^2}$ and $\frac{\partial^2 I}{\partial y^2}$ are the second-order derivatives. The descriptors are further whitened in each part. For color images, we convert the color space to HSV and extract such descriptors in each color channel separately.

2.2 GMM Encoded Commonness

We use a generative model, in particular, a Gaussian Mixture Model (GMM) to capture the commonness of shape and appearance of human bodies. Given the training images, we learn a GMM for each part using the extracted local descriptors. The learned model is denoted by $\Theta = \{(\mu_k, \sigma_k, \pi_k) : k = 1, \ldots, K\}$, where μ_k, σ_k and π_k are the mean, covariance and prior probability of the k-th Gaussian component, respectively. In our implementation, K is empirically set to 16 for each body part and σ_k is diagonal [9].

Since we train the GMM with various images of different people, we consider the trained GMM a representation of people commonness. In essence, these learned Gaussian components can be regarded as the building blocks of the shape and appearance of human bodies while the prior distribution $\Pi = [\pi_1, \ldots, \pi_K]^\top$ describes how the common human body is constructed by these building blocks. We can compute the posterior probability w_{ik} of a local descriptor f_i being generated by the kth Gaussian component

$$w_{ik} = p(k|f_i; \mu_k, \sigma_k) = \frac{\pi_k \mathcal{N}(f_i; \mu_k, \sigma_k)}{\sum_{j=1}^{K} \pi_k \mathcal{N}(f_i; \mu_k, \sigma_k)}. \tag{3}$$

The distance function $D_{com}(\cdot)$ is then defined as

$$D_{com}(L, I_s, T) = \| \sum_{i \in idx(I_s(L))} ([w_{i1}, \ldots, w_{iK}]^\top - \Pi) \|_2, \tag{4}$$

where $idx(I_s(L))$ is the set of indices within the window $I_s(L)$. Intuitively, if the posterior of different Gaussian components deviates far from the prior mixture weights, the contents of the window are less likely to resemble a human body. We use this measure to eliminate false positives that do not contain or well localize a human body.

2.3 Fisher Vector Encoded Uniqueness

The Fisher vector [11] is an image representation which is usually used in visual classification and has seen success in person re-identification. Hereby, we incorporate it with the learned GMM model to describe the uniqueness of a specific person's appearance. For a given window W, we compute the mean and covariance deviation vectors u_k and v_k for each of the K components in the Gaussian mixture models, and the Fisher vector is the concatenation of the deviation vectors u_k and v_k, i.e., $\Phi(W) = [\, u_1, v_1, \ldots, u_K, v_K\,]^\top$, where

$$\begin{aligned} u_k &= \frac{1}{|W|\sqrt{\pi_k}} \sum_{i \in idx(W)} w_{ik} \frac{f_i - \mu_k}{\sigma_k}, \\ v_k &= \frac{1}{|W|\sqrt{2\pi_k}} \sum_{i \in idx(W)} w_{ik} \left[\left(\frac{f_i - \mu_k}{\sigma_k} \right)^2 - 1 \right]. \end{aligned} \tag{5}$$

The distance function $D_{uni}(\cdot)$ is then defined as

$$D_{uni}(L, I_q, I_s) = \| \Phi(I_s(L)) - \Phi(I_q) \|_2 . \tag{6}$$

The final decision of whether a candidate window contains the queried person is made by jointly measuring the commonness and uniqueness (Eqn. 1) with a proper threshold learned from the training data.

2.4 Searching Strategy

The entire framework follows the sliding window strategy. For each candidate window, a binary decision of whether it includes the queried person is made. We further perform multi-scale scanning against scale variations and non-maximum suppression to remove redundant responses. One different strategy employed in our framework is that the redundant responses are pruned using additional identity information, that is, if two overlapped windows are identified as two different subjects, they will not be merged; while in the original protocol for detection, responses of all people are put together for pruning and hence occluded people can be incorrectly merged with the person in the front. This way we can effectively reduce false negatives.

3. EXPERIMENTS

In this section, we show our experimental results and comparisons to other approaches on public benchmarks.

3.1 Datasets and Settings

We validate our method on two public datasets: CAMPUS-Human [12] and EPFL [5], as shown in Tab. 1. The ground truth ID and bounding box of each target person are provided in these two datasets. Other public Re-ID datasets are unsuitable for our task without scene images.

Table 1: Dataset information

Dataset	ID	Query	Scene	Target	Scene Res.
CAMPUS	74	370	214	1519	640×360
EPFL	30	70	80	294	360×288

To better evaluate our method, we separate the CAMPUS-Human dataset into two different scenarios, one crowded scenario S1 with more than 10 people in one shot and one sparse scenario S2 with less than 10 people. For both datasets, we use all query images belonging to the same person for search, and scanning every scene shot to locate the queried person. All the parameters are fixed in the experiments, including scanning scale factor 1.1, window size 60×180, window stride 8 (on both horizontal and vertical directions).

3.2 Experimental Results

To objectively evaluate our approach, we setup two baselines for comparison: 1) sequentially applying HOG-based people detection [1] and person re-identification using Fisher vectors [9] (*Detect+ID Seq*); 2) combining these two techniques more tightly by fusing the detection score and the identification difference (*Detect+ID Jnt*), the fusion weight is found by cross validation and set as 0.25(4−DetectScore)+ IDDifference. Besides, we analyze the component effectiveness of the proposed method with two settings: 1) using only uniqueness as in the general image search task (*Our Uniq*); 2) using both commonness and uniqueness (*Our Full*).

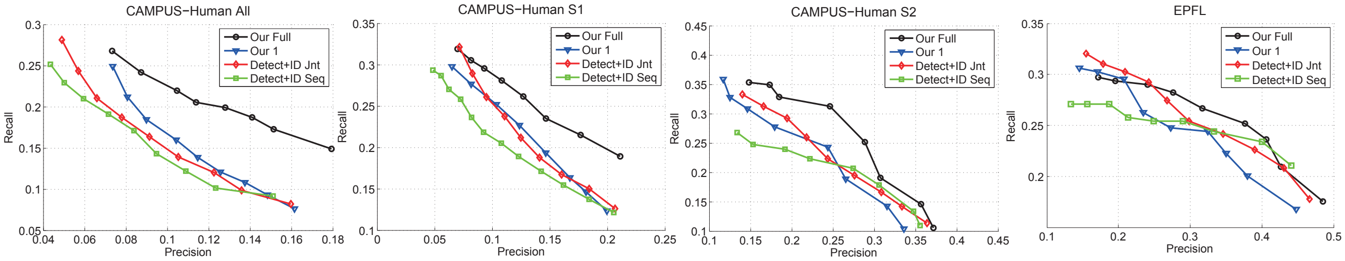

Figure 2: Performance comparisons using PR curves on CAMPUS-Human All, CAMPUS-Human S1, CAMPUS-Human S2 and EPFL datasets.

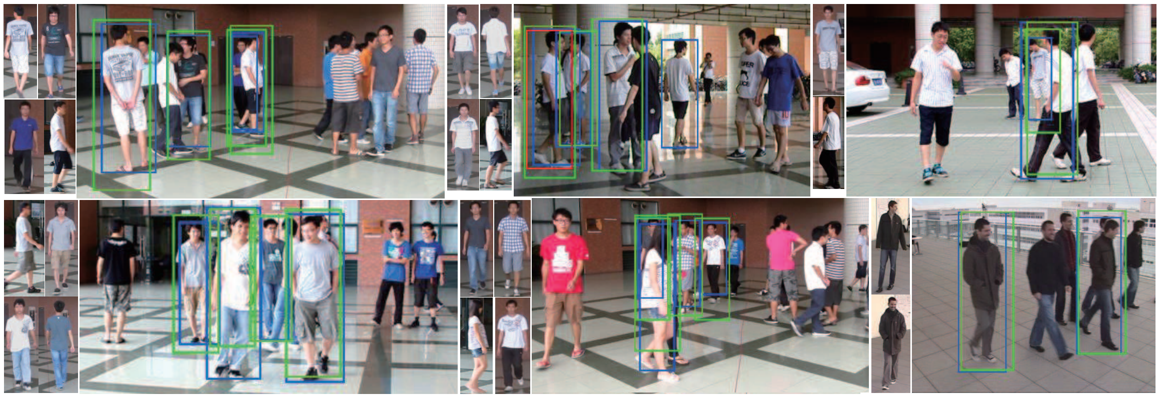

Figure 3: Person search results generated by our method. Given query persons, green and red rectangles in scene images are respective correct and false locations, compared with the groundtruth with blue ones.

Table 2: Quantitative results on two public datasets.

Method	CAMPUS-Human			EPFL
	All	S1	S2	
Our Full	**16.28%**	**19.96%**	**27.55%**	**30.15%**
Our Uniq	13.59%	17.65%	24.47%	27.87%
Detect+ID Jnt	12.15%	16.54%	24.12%	29.72%
Detect+ID Seq	11.51%	15.82%	23.96%	29.54%

We adopt the PASCAL Challenge criterion to evaluate the localization results: a match is counted as the correct match only if the intersection-over-union ratio (**IoU**) with the groundtruth bounding box is greater than 50%. We utilize the F-score as the benchmark metric, which measures the accuracy of person search by considering both recall and precision. The F-scores and PR curves of all experiments on both datasets are quantitatively reported in Table. 2 and Fig. 2, and qualitatively shown in Fig. 3, respectively.

From the results, we conclude that our unified framework outperforms the other searching strategies in general. *Detect+ID Jnt* works better than *Detect+ID Seq*, supporting our advocation for joint modeling of detection and identification. However, because the detection score and identification score are two naturally different metrics, the direct combination of such scores does not work as well as our fundamentally coupled framework (*Our Full*). The improvements of *Our Full* over the baselines are smaller on the EPFL dataset because this dataset is relatively easy for the detectors.

4. CONCLUSIONS

In this paper, we propose a unified framework for person search, a practical multimedia search problem, by jointly modeling the commonness of people and the uniqueness of the queried person. The efficiency is still a concern due to the high computational cost in each sliding window. We are exploring the state-of-the-art region proposal methods to replace the sliding window protocol, as well as more efficient approaches to model commonness and uniqueness.

5. REFERENCES

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *CVPR*, 2005.

[2] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *TPAMI*, 2012.

[3] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. *CVPR*, 2010.

[4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *TPAMI*, 2010.

[5] F. Fleuret, E. Turetken, and P. Fua. Multiple object tracking using k-shortest paths optimization. *TPAMI*, 2011.

[6] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. *ECCV*, 2008.

[7] P. Jain, B. Kulis, and K. Grauman. Fast image search for learned metrics. *CVPR*, 2008.

[8] W. Li and X. Wang. Locally aligned feature transforms across views. *CVPR*, 2013.

[9] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. *ECCV Workshops*, 2012.

[10] Y. Rui, T. S. Huang, and S.-F. Chang. Image retrieval: Current techniques, promising directions, and open issues. *J. Vis. Commun. Image Represent.*, 10(1):39–62, 1999.

[11] J. Sanchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *IJCV*, 2013.

[12] Y. Xu, L. Lin, W. Zheng, and X. Liu. Human re-identification by matching compositional template with cluster sampling. *ICCV*, 2013.

[13] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. *CVPR*, 2013.

[14] W. Zheng, S. Gong, and T. Xiang. Re-identification by relative distance comparison. *TPAMI*, 2013.