PeR-ViS: Person Retrieval in Video Surveillance using Semantic Description

Parshwa Shah
School of Engineering and Applied Science, Ahmedabad University
parsha.s.btech16@ahduni.edu.in

Arpit Garg, Vandit Gajjar
School of Computer Science, The University of Adelaide
{arpit.garg, vanditjyotindra.gajjar}@student.adelaide.edu.au

Abstract

A person is usually characterized by descriptors like age, gender, height, cloth type, pattern, color, etc. Such descriptors are known as attributes and/or soft-biometrics. They link the semantic gap between a person’s description and retrieval in video surveillance. Retrieving a specific person with the query of semantic description has an important application in video surveillance. Using computer vision to fully automate the person retrieval task has been gathering interest within the research community. However, the current trend mainly focuses on retrieving persons with image-based queries, which have major limitations for practical usage. Instead of using an image query, in this paper, we study the problem of person retrieval in video surveillance with a semantic description. To solve this problem, we develop a deep learning-based cascade filtering approach (PeR-ViS), which uses Mask R-CNN [14] (person detection and instance segmentation) and DenseNet-161 [16] (soft-biometric classification). On the standard person retrieval dataset of SoftBioSearch [6], we achieve 0.566 Average IoU and 0.792 %IoU > 0.4, surpassing the current state-of-the-art by a large margin. We hope our simple, reproducible, and effective approach will help ease future research in the domain of person retrieval in video surveillance. The source code will be released after the paper is accepted for publication with baseline and pretrained weights. The source code and pretrained weights available at https://parshwa1999.github.io/PeR-ViS/.

1. Introduction

Recently, pedestrian attribute recognition such as age, gender, height, cloth color, and type, etc. has obtained increasing attention due to its promising outcomes in applications such as person re-identification, attribute-based person search, and person retrieval in video surveillance. Nowadays metropolitan cities are equipped with thousands of surveillance cameras, which stores a gigantic amount of surveillance data every second. To retrieve a specific person manually from large-scale videos possibly takes months to complete. Using computer vision techniques to fully automate the above task has been gathering interest within the research community. The current trend mainly solves this task based on image queries, which have major limitations and might not be suitable for practical usage. Therefore, we studied the problem of person retrieval with semantic descriptions to face these limitations. Figure 1 illustrates the example of the person retrieval using a semantic description.

Figure 1: Person retrieval using a semantic description. Given the semantic description of a person, our approach PeR-ViS uses each description in a filtering mode to retrieve the correct person.

Person retrieval with image-based queries is known as Person Search and/or Person Re-identification in computer vision [38, 20, 33]. Given image query to network, it finds the similarity between the query and that surveillance footage. The most identical person than retrieved from the footage according to similarity score. However, this prob-
lem requires at least one image as a query for the network, which has a major limitation in practice. In many cases, such as a lost person, there might be only a description provided of the person(s) appearance available. Facing the limitation of image-based person retrieval, we propose to use a semantic description for person retrieval. It does not require a person(s) image to be applied to the network. The semantic description can also accurately describe the information of the person(s) appearance.

To use this semantic description for person retrieval, we propose PeR-ViS - a deep learning-based cascade filtering approach for person retrieval in video surveillance. The PeR-ViS takes a semantic description and a surveillance frame as an input and outputs the correct retrieved person with a bounding box. It uses Mask R-CNN [14] for precise detection and instance segmentation of every person in the surveillance frame. It uses Height, Torso Module (Color >> Type >> Pattern), Leg Module (Color >> Pattern) and Gender as cascade filter. These descriptors are chosen due to their distinguishable capability. For example, the height descriptor is view and distance invariant, while predicting color is also invariant to angle and direction [27]. The height filter is designed using camera calibration parameters, while all other filters are based on convolutional neural networks: DenseNet-161 [16]. The complete approach is illustrated in Figure 2, and we will talk about more in Section 3.

In summary, the main contributions of this paper are four-fold:

• We study the problem of person retrieval with the semantic description in video surveillance, which often arises in real-life scenarios, but remains wide open in the research community.

• Mask R-CNN is used by us because of following advantages
  – Unnecessary background clutter is removed.
  – As accurate segmented boundary provides precise head and feet points. Better estimation of real-world height can be made.
  – It helps to extract accurate patch for Torso and Leg module attribute classification.

• The estimated height can also be used to distinguish between the standing and sitting position of the person. This ability narrows down search space for the person of interest in the standing position.

• We propose a new person retrieval approach (PeR-ViS) which uses cascade filters of person(s) descriptors to narrow down the search space of detected people to leave only the target. The approach surpasses the current state-of-the-art on the SoftBioSearch dataset, achieving 0.566 Average IoU and 0.792 %w IoU > 0.4.

The rest of this paper is as follows. Section 2 describes work-related to person search and retrieval in video surveillance. Our approach to PeR-ViS and its modules is briefly mentioned in Section 3. The experiment, the implementation details, and the results are described and shown in Section 4. Section 5 discusses the experimental results and ablation studies. Finally, Section 6 focuses on the possible future work and concludes the paper.

2. Related work

Person Search: A person search is a recently introduced problem, where an image query is applied to the network and the same/similar person can be found. Li et al. [34] have proposed a person search task that aims to find a similar person(s) in the photo-gallery without bounding box annotation. The respective data is similar to that in the person re-identification task.

Person re-identification: Person search is indeed an extension of the Person re-identification task [36, 10], which objective is to match a person(s) image from various cameras within a short span. The problem has drawn much attention in the computer vision research community since the last decade. Several datasets have been proposed to tackle the task of person re-identification. Early person re-identification methods focus on manually designing distinguishable features [13, 31], and learning distance metrics [19, 21, 24]. With the growth of deep learning in recent years, many researchers have proposed several deep learning-based solutions to solve this task. Li et al. [19] and Ahmed et al. [2] created CNN models for person re-identification. Both the CNNs uses a pair of cropped person(s) images as input and utilizes a binary verification loss function to train the different parameters. Ding et al. [7] and Cheng et al. [3] utilized triplet instances for training CNNs to minimize the feature distance between the similar person and maximize the distance between different people. Instead of utilizing the triplet loss function, Xiao et al. [32] proposed to learn features by categorizing identities.

Person Retrieval: Before the growth of deep learning in the field of computer vision, hand-crafted methods had developed algorithms to learn part or local features. In the work of Gray et al. [12], they have partitioned
3. Proposed approach - PeR-ViS

This section introduces a deep learning-based cascade filtering approach for person retrieval in video surveillance (PeR-ViS). Figure 2 illustrates the complete flow diagram of PeR-ViS. Each video surveillance frame is given to state-of-the-art Mask R-CNN [14] in order to detect and instantiate segment(s) of person(s). For person(s) head and feet, points are extracted for all detected and segmented person(s). For height estimation, using the camera calibration technique, it is calculated based on the height given in the semantic description. In this complete approach, height acts as a primary filter to narrow down the search space of person(s) in the frame. In the case of multiple matches, where more than one person matches the height description, additional filtering is performed using torso color, type, pattern; leg color, pattern, and gender. Instance segmentation helps us in obtaining background free extraction of the patch from the torso and leg. The number of identities is further narrowed down by comparing the semantic description with the extracted patches. The preciseness of the final output is further improved by exploiting gender classification. The following sections describe the process of filter modules for person retrieval.

**Height estimation:** Person height is view-invariant, which helps to distinguish between standing and sitting position of the person. Tsai camera calibration approach [29] is used to estimate detected person(s) height by matching into horizontal stripes to extract color and texture features. Similar partition has also been used by many other works [20, 25, 23]. We also have adopted the strategy for extracting torso and leg patches accurately. Some other works utilize a more refined strategy. In the work of Gheissari et al. [10], authors have divided person(s) into different triangular parts for feature extraction. Cheng et al. [4] used a gentle structure to parse the person(s) into different semantic parts. Das et al. [5] have applied histograms on the head, torso, and leg portion to extract the spatial information. The state-of-the-art on most person retrieval datasets is currently maintained by deep learning algorithms [37]. There are other methods, which are more or less similar to our work for person retrieval, however, the strategy of cascade filtering that aims to narrow down the detected people to leave only the target is our major contribution.
bounding box coordinates to real-world coordinates. The dataset of SoftBioSearch [6] provided 6 calibrated cameras for the calculation of real-world coordinates. Detected person(s) head and feet points are computed from the instance segmentation, which can be seen in the height filter (Figure 1). Steps for Computation of height estimation are as follows:

- From the given camera calibration parameters Intrinsic parameters matrix \(I_m\), rotation matrix \(R_m\) and a translation vector \(t_v\) are computed.
- The respective transformation matrix is computed as \(T = I_m[R_m|t_v]\).
- By using radial distortion parameters, head and feet points are undistorted.
- Using inverse transformation of \(T\) global coordinate for feet is set to \(F = 0\) and global coordinates \(X, Y\) are derived.
- By using \(X, Y\) coordinate to compute the \(F\) coordinate of the head which also characterizes height.

Calculated height assists in narrowing down the search space within the test surveillance frame based on the semantic description (e.g. Average height (150-170 cm)). After that test surveillance frame now only contains the person(s) which matches the height’s semantic description. During training time, annotated head and feet points are used to calculate the height from all the input surveillance frames of the video sequence. The average height \(H_{avg}\) is computed over all the surveillance frames in a given video sequence. Over the similar training video sequence, we noticed that the average height estimated from automated head and feet point is greater than was \(H_{avg}\). This dissimilarity holds the wrong computation, therefore; it is subtracted from \(H_{avg}\) during the testing time to equalize the error.

![Figure 3: Extraction of torso and leg region from detected person(s) body.](image)

**Torso and Leg Module Filter:** Mask R-CNN generates a person(s) label with a confidence score and instance segmentation. Using the golden ratio for height, the torso and leg regions are extracted. From the detected bounding box and instance segmentation, the upper 20-50% region is classified as the torso, while the 50-70% represents the leg portion of the person(s). The extraction of these regions is shown in Figure 3. Instance segmentation is utilized to get the torso and leg portion without the background to refine the cloth color, pattern, and type classification. Patch for prediction is extracted from a region marked by points ‘a’, ‘b’, ‘c’, and ‘d’ (ref. Figure 1 - Torso and Leg module filter). The extracted patches are used to fine-tune DenseNet - 161 [16], for predicting the confidence score. The SoftBioSearch dataset [6] annotation contains 13 Torso primary and secondary color, 4 torso cloth type, 8 torso cloth pattern, 13 Leg cloth primary color and secondary color, and 8 leg cloth patterns for each identity. In the case of different similar matches, the approach will refine the results by using each given modality. This feature of cascade filtering helps to narrow down the search space. Furthermore, to get and/or verify the result of the last filter (Leg cloth pattern), we use the gender filter. We noticed that in some of the cases gender filter helps to improve the performance when the surveillance frame is crowded.

**Gender Classification:** By using height and other semantic descriptions in a cascade filtering mode, the proposed approach retrieves the person for most cases. But when multiple matches come after the last filter, the approach uses Gender as the final filter for retrieving or verifying the result. Full-body images of male and female categories were used in fine-tuning DenseNet.

4. Experiments

This section discusses the details about the overview of the dataset, performance metric for evaluation, and implementation details. The PeR-ViS uses Mask R-CNN, which uses the pre-trained weights of Microsoft COCO [22] dataset for detection and instance segmentation. The Mask R-CNN model uses the ResNet-101 FPN as the backbone architecture, which has achieved Average Precision (AP) of 35.7 on the COCO test-dev set.

4.1. Dataset overview

Our work uses the SoftBioSearch database [6], which uses 6 stationary calibrated cameras and consists of 110 unconstrained training video segments. Each of these 6 cameras is annotated with Tsai’s camera calibration technique [29]. Every video sequence in the training set is labeled with a set of semantic descriptions to describe the person(s) identity. In addition, with the provided description, nine key-points were also annotated in the case to use human key-point estimation and exploits with the person retrieval task. More precise details on the camera calibration and human body key-point can be found in [6].
The test set consists of further 41 person identities taken from 4 of the 6 cameras (Here camera numbers 1 and 6 are not used) used in collecting the training set. Compilation and annotation of the test set follow the same instructions of the train set to ensure parity, with at least the first 30 surveillance frames of each video sequence reserves to allow the person(s) identity to fully enter in the camera view.

For more details on the performance evaluation, the test person identities were separated into very easy, easy, medium, and hard.

The given labels are defined below:

- **Very Easy**: randomly populated scenario, no complex factors, and target person identity clearly visible.

- **Easy**: Scenario consists of one or more people, but the person(s) identity is clearly distinguishable.

- **Medium**: From the following factors one of the factors will be present: Similar type of the identity present in the scenario, Very heavy occlusion with the target, a crowded scene.

- **Hard**: Two or more of the above factors present in the scenario. In the test set out of 41 person identities, 6 are labeled as very easy, 13 as easy, 12 as the medium, and 10 as hard.

### 4.2. Performance metric

As described in SoftBioSearch dataset, the metrics use the Intersection Over Union (IoU) given by Eq. 1. An IoU$_{avg}$ is calculated per video sequence, and video sequences results are averaged over all video sequences to obtain a final accuracy measure. IoU = $\frac{\text{D} \cap \text{GT}}{\text{D} \cup \text{GT}}$. Where D is the obtained bounding-box from the approach and GT is the Ground Truth bounding box.

### 4.3. Implementation details

The fine-tuning is accomplished on a workstation with an Intel Xeon core processor and accelerated by NVIDIA TitanX 12 GB GPU. All experiments run in Tensorflow 1.8.80-20% train and validation set.

**Data Augmentation**: In order to achieve generalization of training data for improved performance and robustness data augmentation is used. E.g. In the training set after removing the surveillance frames with partial occlusion, the final set contains 8577 images from 110 subjects. Thus training DenseNet with only 8577 images may result in over-fitting, which is avoided using a data augmentation scheme. Each training frame is horizontally and vertically flipped, rotated with 10 angles $1^\circ, 2^\circ, 3^\circ, 4^\circ, 5^\circ, -1^\circ, -2^\circ, -3^\circ, -4^\circ, -5^\circ$ and brightness increased with a gamma factor of 1.5.

**DenseNet training for Cloth Color, type, pattern; and gender**: Cloth Color, Type, Pattern, and gender models are fine-tuned using DenseNet - 161 which is pre-trained on the ImageNet dataset. The SoftBioSearch dataset consists of 1704 patches divided into 13 torso and leg primary and secondary colors; 8 torso and leg patterns; and 4 torso type. Extra patches for training these attributes are extracted using 4 human key-points provided in annotations (Left-right shoulder and left-right waist). In order to deal with light changes, these patches are augmented by increasing the brightness with a gamma factor of 1.5. Thereafter, almost 17000 patches are generated and further divided into 80-20% train and validation set.

All the networks for the torso and leg module are trained using mini-batch stochastic gradient descent (SGD). Due to the computation cost, the fine-tuning strategy is the same for all the descriptors. The networks were trained for 20 epochs with a learning rate is set to 0.001, dropout set to 0.35, and effective mini-batch size is set to 32. Table 1 shows the validation accuracy of different descriptors.

For the gender classification, the initial data augmentation generated 105980 images for training gender descriptor, which is almost 13 times larger than the original training set (8577). 20% of total images were used for validation. For the gender fine-tuning, the network was trained for 30 epochs with a learning rate of 0.01, dropout rate of 0.25, and effective batch size is set to 64.

### 5. Experimental results

This section covers the qualitative and quantitative experimental results derived from a test set of 41 person identities. Overall results are shown in Table 2, including a comparison with the baseline algorithm and current state-of-the-art methods. Form this we can see that our approach - PeR-ViS outperformed the current state-of-the-art by a large margin.

Considering the algorithms used, [35], [9], and [26] uses deep learning method that deploys a CNNs for detecting person(s) identity in the surveillance frame. The baseline of [6] uses an avatar, which is a non-deep learning approach, that is constructed from the input semantic query to drive

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Validation Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Torso Color</td>
<td>81.29%</td>
</tr>
<tr>
<td>Torso Type</td>
<td>79.14%</td>
</tr>
<tr>
<td>Torso Pattern</td>
<td>76.5%</td>
</tr>
<tr>
<td>Leg Color</td>
<td>71.52%</td>
</tr>
<tr>
<td>Leg Pattern</td>
<td>72.5%</td>
</tr>
<tr>
<td>Gender</td>
<td>77.79%</td>
</tr>
</tbody>
</table>

Table 1: **Accuracy results** on the validation set of different semantic descriptor based on the above hyper-parameter setting.
Approach & Average IoU & %w $\text{IoU} > 0.4$
\hline
Baseline [6] & 0.290 & 0.669 \\
Galiyawala et al. [9] & 0.363 & 0.522 \\
Schumann et al. [26] & 0.503 & 0.759 \\
Yaguchi et al. [35] & 0.418 & - \\
Yaguchi et al. [35] & 0.462 & - \\
Yaguchi et al. [35] & 0.511 & 0.669 \\
Ours & 0.566 & 0.792 \\
\hline

Table 2: Overall IoU of different methods on the test set.

![Graph showing performance broken down by sequence difficulty and compared with Average IoU.](image1)

![Graph showing performance broken down by sequence difficulty and compared with %w IoU > 0.4.](image2)

Figure 4: Performance broken down. (a) by sequence difficulty and compared with Average IoU. (b) by sequence difficulty and compared with %w $\text{IoU} > 0.4$. (from top to bottom - a, b)

A gradient descent search. It is noticeable that both [9] and [6] uses very few descriptors, while [35] and [26] uses the full set of available descriptors, likely improving the performance.

To further analyze performance, we break the results by difficulty level (Figure 4) described in Section 4.1 and into individual video sequences (Figure 5). From Figure 4 it is clear performance decrease with an increase in difficulty. The “Very Easy” video sequences mainly contain only a single moving person identity, and as seen in the figure all other methods perform well on these video sequences. As the difficulty level increases, [9] suffers a very much decrease in the performance compared to others. A very slight difference can be spotted in the performance of [35] and [26]. Both achieve very similar IoU’s, but as compared with our approach their algorithms fail to perform well in the hard video sequences. Thus indeed our approach performs very well on medium and hard sequences and breaks the current state-of-the-art by a margin of more than 5%. Figure 5 shows the performance of PeR-ViS on each video sequence. It is noticeable that performance varies across each video sequence and several sequences pose a challenge for retrieval. From Figure 4 as compared with [26], the authors use a tracking approach, which helps to reduce the error that may impact non-tracking approaches such as ours and others. From this, it is noticeable that ours and [35] have the inclination to either detect a person’s identity very precisely or very badly.

Surprisingly, two of the video sequences (20 and 39) contain very less crowd, but very complicated by having a very similar person identity. The other two video sequences consist high amount of crowd, and we see almost all the systems are having a decrease in performance. To further showcase the person retrieval results, Figure 6 and Figure 7 show some of the examples of True positive and False negative cases based on semantic descriptions.

Figure 6 shows true positive cases of person retrieval using semantic descriptions. In Figure 6, images from left to right indicate the input test frame, the output of the height module filter, the output of Torso and Leg module filter and gender module filter respectively. Figure 7 shows the results when the approach fails to retrieve the person correctly. It also indicates that the dataset is created in challenging conditions. In Figure 7, the approach fails due to following results: (a) Multiple persons with the same torso color yields incorrect color classification, (b) Multiple person(s) with occlusion, (c) Same height class appears when multiple persons comes in the surveillance frame and (d) Mask R-CNN fails for person detection.

5.1. Ablation experiments

We run some ablations to analyze our PeR-ViS approach

**Choice of classification network:** To test the influence of a classification network on the proposed approach, we have formed five video sequences from test set consists of Easy, Medium, and Hard category. We have used AlexNet, VGG-16, and ResNet-152 for ablation experiment and average IoU is used for evaluation. All the networks perform well on the easy video sequences. Next on the medium sequences, where background clutter and less occlusion are present, AlexNet and VGG-16 perform poorly. Here ResNet-152 shows equivalent performance to DenseNet-161. The true performance of DenseNet-161 is noticed in the hard category, where it achieves excellent performance. The Table 3 shows the result of IoU using different networks.

**More descriptors:** In our work, we have used Height; Torso and Leg module; and Gender descriptors. The choice of these descriptors is purely based on perceptive ability. We have tried to add more descriptors such as Age, Shoe color, Luggage, and Human body’s build but none of the modality was capable to improve the performance reported.
Figure 5: **Per Sequence Performance** for test set.

(a) Person retrieved using only height.

(b) Person retrieved using height and torso module filter.

(c) Person retrieved using height, torso, leg and gender module filter.

Figure 6: **True positive cases** of person retrieval with semantic description. (Best viewed in color and magnification.)
Figure 7: **False negative cases of person retrieval.** (a) incorrect color classification with multiple person(s), (b) multiple person(s) with occlusion, (c) multiple person(s) with same torso color and height class and (d) person detection fails. The supplementary material contains more true positive results. (from left to right – a, b, c, d) (Best viewed in color and magnification.)

<table>
<thead>
<tr>
<th>Network</th>
<th>Very Easy</th>
<th>Easy</th>
<th>Easy</th>
<th>Medium</th>
<th>Hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlexNet</td>
<td>0.567</td>
<td>0.534</td>
<td>0.523</td>
<td>0.325</td>
<td>0.183</td>
</tr>
<tr>
<td>VGG-16</td>
<td>0.641</td>
<td>0.621</td>
<td>0.615</td>
<td>0.336</td>
<td>0.237</td>
</tr>
<tr>
<td>ResNet-152</td>
<td>0.742</td>
<td>0.712</td>
<td>0.64</td>
<td>0.492</td>
<td>0.36</td>
</tr>
<tr>
<td>DenseNet-161</td>
<td>0.762</td>
<td>0.733</td>
<td>0.642</td>
<td>0.582</td>
<td>0.461</td>
</tr>
</tbody>
</table>

Table 3: **This table shows the IoU result** on 5 video sequences (Sequence Number 4, 13, 21, 23, and 28) using different network architecture in our approach.

### 6. Discussion and Conclusion

The proposed approach - PeR-ViS retrieves the person in video surveillance based on the semantic description of Height; Torso Cloth color, type, and pattern; Leg Cloth color and pattern; and Gender. The major benefit from this filtering sequence is that Height, and Torso Color, Type, and Pattern are easily differentiable compared to other descriptors, where the heavy crowd is present, leg patch won’t easily extract. Thus the choice of this filtering sequence is most important in our work. Also, instance segmentation allows precise height estimation and accurate color patch extraction from the torso and leg. Thus, our algorithm achieves an average IoU of 0.566 and %w IoU > 0.4 of 0.792, surpassing the current state-of-the-art by a large margin. We have also provided the code snippet up-to Torso type filter (Height >> Torso Cloth Color >> Torso Cloth Type and Gender for verification). The possible future work will now focus on how to improve the results by incorporating human pose estimation and other descriptors and investigate the architecture for generalization in real-world scenarios. Furthermore, the tracking approach is also useful when the person is retrieved with Height and/or Torso module filter, which will essentially lower the computation usage.

### Acknowledgements

We would like to thank anonymous reviewers for providing us their valuable feedback on our paper. We would like to express our deep gratitude to Dr. Mehul S. Raval, Mr. Hiren Galiyawala and Mr. Kenil Shah for providing useful comments and discussion. We would also like to thank Ms. Ayesha Gurani, Mr. Viraj Mavani and Mr. Yash Khandhediya for their help with manuscript.

### References


[31] Xiaogang Wang, Gianfranco Doretto, Thomas Sebastian, Jens Rittscher, and Peter Tu. Shape and appearance con-


