An Overview of Recent Approaches in Person Re-Identification

Abstract
Deploying large number of cameras with overlapping views for surveillance in large areas such as airports, shopping malls, and elderly care centers is not cost effective. Tracking people using computer vision algorithms in such scenarios involves matching people in non-overlapping camera views. This problem is referred to as Person Re-Identification in the computer vision literature. The important challenges of this problem are illumination variation, pose variation and low image resolution. Recently, Person Re-Identification has gained increased interest among researchers and a plethora of potential solutions is presented in the literature. In this paper an overview of recent approaches in Person Re-Identification is provided.

1. Introduction
Variations in the person’s appearance within a short period of time arise due to changes in pose, illumination, expression, and occlusion. Over a long period of time variation arises naturally due to aging. Biometric solutions like Face recognition, Gait recognition, Ear recognition, etc. has been proposed in the literature to solve both the short-time interval and long-time interval problems together. The recently investigated Person Re-identification problem focuses only on the variations that arise during a short period of time. The important application of Person Re-Identification is to associate people in multi-camera surveillance systems across non-overlapping camera views at different locations. The important characteristics of this problem are the changes in the view point (view angle and distance), occlusion, lighting condition, camera parameters and background across different camera views.

![Figure 1 Example probe and gallery of images](image-url)
Though the general re-identification problem involves matching people in the videos captured across camera networks, there are critical assumptions in the solutions provided in the literature. One assumption is that the bounding box for the people in all the images in all the videos are available. This reduces the problem from matching people in videos to matching the given image regions. Another assumption is that the appearance of the person does not change, meaning the person does not change the clothing during the search time. Because of this assumption the solutions in the literature models the appearance of the clothing, instead of the person.

The person who needs to be searched across the camera network is provided as a cropped image region and is called the probe image. The videos that need to be searched are represented by the cropped image regions of the people. This set of images to be searched are called the gallery images. Re-Identification problem is defined as ranking the gallery images for a given probe image. An example probe and gallery image is shown in Figure 1. Performance of the algorithms is compared based on their effectiveness in ranking the correct gallery image within the top few ranks. In the literature cumulative matching characteristic (CMC) curve is used to compare the performance of the algorithms. CMC curve provides information about the number of gallery images to be searched to find the correct match. The datasets used in the literature can be found in Table 1.

<table>
<thead>
<tr>
<th>Name</th>
<th>Image/Video</th>
<th># people</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIPeR</td>
<td>Images – 1264</td>
<td>632</td>
<td>Outdoor – University</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td># cameras: 2</td>
</tr>
<tr>
<td>iLIDS</td>
<td>Images – 476</td>
<td>119</td>
<td>Indoor – Airport</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td># cameras: 5</td>
</tr>
<tr>
<td>CAVIAR4REID</td>
<td>Images, &gt; 1000</td>
<td>50</td>
<td>Shopping mall</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td># cameras: 2</td>
</tr>
<tr>
<td>ETHZ</td>
<td>Videos</td>
<td>146</td>
<td>Outdoor – Road Moving camera</td>
</tr>
<tr>
<td>3DPeS</td>
<td>Videos</td>
<td>130</td>
<td>Outdoor – University</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td># cameras: 8</td>
</tr>
<tr>
<td>QUML</td>
<td>Images – 500</td>
<td>250</td>
<td>Indoor – Underground station</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td># cameras: 8</td>
</tr>
<tr>
<td>Shinpuhkan</td>
<td>Images – 22000</td>
<td>24</td>
<td>Shopping mall</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td># cameras: 16</td>
</tr>
</tbody>
</table>

Table 1 Some Popular Dataset Details

This paper is organized as follows: the typical computer vision steps involved in solving the re-identification problem is explained in section 2. The typical features and their representation and the distance metrics used in the literature are also provided. Performance measures are discussed in section 3. Some key approaches used in the literature are briefly explained in section 4. Each of these approaches is unique in the intuition or emphasis provided in the solution. The approaches are classified into the following categories: Feature representation [13, 14], Saliency [11], Machine learning [5, 6, 7], Collaborative representation [9], Domain transfer [10], Attributes [8], Brightness transfer function [18], and Group classification [12]. Future work is discussed in section 5.
2. Re-Identification Pipeline:

Re-identification problem in the literature assumes that the bounding box for the person in the image is provided. This avoids the need for having a person detector module. Typical re-identification processing pipeline involves image segmentation, feature extraction, feature representation and matching. The matching stage gives a score. Matching score between the probe image and all the gallery images are computed and is used for ranking the gallery images.

![Re-Identification Pipeline](image1)

2.1. Image segmentation

Segmentation helps to encode the location information of the features. The common ways of segmenting the images are global, horizontal strips and part based. No segmentation operation is performed in the global segmentation method, the extracted features has to encode the location information. For example, the covariance feature with the \( (x, y, R, G, B) \) descriptor for each pixel encodes the pixel local information. The most common segmentation method is the horizontal strips. The horizontal strips based methods vary in terms of the number, size and the overlap ratio of the strips. Some methods segment the images into body parts: head, torso, arms and legs. Arm part is generally small and is ignored by some methods.

![Typical Image segments used for Re-Identification](image2)
2.2. Features and Representations

Re-identification methods make the assumption that the appearance (clothing) of the person does not change. This assumption implies that the color and texture information of the clothing can be used for identification. The typical color features used are histograms of the channels of different color spaces, namely, RGB, HSV, YCbCr, LAB and Log-Chromaticity. Some of the other color features are dominant colors and global color context. Dominant colors are obtained by clustering the pixel colors channels values and selecting the top few clusters. In global color context some prominent colors such as red, blue, green, yellow, orange, purple, etc are defined. Then a histogram is created for the number of pixels close to each of these prominent defined colors.

The common texture features used are Histogram of Oriented Gradients (HOG), Scale-Invariant Feature Transform (SIFT), Speeded Up Robust Features (SURF), PHOG, Haar-like, Local Binary Patterns (LBP), Color Co-occurrence, Gabor and Schmid filters.

Some of the features, such as HOG, have their own representations, other features like color channels do not have specific representation. Some common feature representations used in the literature are histogram, covariance matrix, Fisher vector, etc.

![Figure 4 Texture Feature - Histogram of Oriented Gradients (HOG)](image)

![Figure 5 Texture feature - Local Binary Patterns (LBP)](image)

![Figure 6 Texture features - Features used to form the covariance matrix](image)
2.3. Distance measures
The commonly used distance measures for the histogram representation are L1 norm, Euclidean, Correlation, Chi-Square, Bhattacharyya and Histogram Intersection. For the matrix representation Generalized Eigen value based distance measures are used. Existing distance measures cannot be used for some custom feature representations, so these approaches use metric learning methods to learn the distance measure.

3. Performance Measure
Cumulative matching characteristic (CMC) curve is the most commonly used performance measure. CMC curve gives the probability that the correct match for a given probe is observed within the top K ranks. A sample CMC curve for the single probe vs single gallery is shown in Figure 9. It is reasonable to expect that multiple images will improve the re-identification performance. Some approaches have exploited the presence of multiple images for the probe and gallery person images. Improvement in the results using multiple images can be observed in the sample CMC curve for the multiple probe vs multiple gallery is shown in Figure 10.
4. Approaches

Several different solutions were proposed in the literature to solve the re-identification problem. In this paper, the approaches are classified into eight categories. Each approach in the literature belongs to one or more of these eight categories. These eight categories are formed based on their novelty and the intuition behind the solution. For example, the intuition behind the solution that learns the features using machine learning method is different from the solution that learns the brightness transfer function between camera pairs.

1. Feature representation [13, 14]
   These approaches differ in the set of features used and their representation. The distance measure can be found for each image segment or feature separately and then combined using weights. The weights for the features can be found experimentally or can be learned.

   These approaches emphasize that features should be weighted for each person differently or the features that are unique to the person should be weighted more than the others.

3. Machine learning [5, 6, 7]
   Machine learning approaches can be used in different ways:
   a. To learn the feature representation instead of hard coding the features.
   b. To find a boundary that classifies the similar and dissimilar pairs of images.
   c. There is no available metrics for different feature combinations. So metrics can be for the specific feature representations can be learned.
4. Collaborative representation [9]
   An images can be represented based on its similarity to the other images in the dictionary. The dictionary captures potential variations in the appearance that can be found in a given scenario. For example, the dictionary can contain people with backpack, beard, various color combination of shirts and pants, etc.

5. Domain transfer [10]
   The number of camera pairs is a quadratic function of the number of cameras. Applying machine learning algorithm for each of these camera pairs is time consuming. Domain transfer approaches try to find a solution to transfer the function learned for camera pair A to camera pair B with little overhead to no overhead, so that complete learning of a new function for camera pair B is avoided. It involves identifying the correct camera pair to transfer and the transfer function.

6. Attributes [8]
   These approaches represent people using a semantic description, instead of using low-level features. Different attributes are defined such as cap, shorts, bag, backpack, shirt, skirt, etc. People are represented as the vector of probability values for the presence of each of these attributes. The probability of each attribute can be computed using a machine learning classifier.

7. Brightness transfer function [18]
   These approaches model the function for the illumination change between a camera pair.

8. Group classification [12]
   In this approach a group of people are identified instead of any single person. Classifying group is more challenging and there are different problems associated with the group classification that does not exist in the re-identifying a single person. People move with in a group causing change in color distribution for any given segment and people get occluded by other members within the group.

5. Future work
   There are some significant difference between the re-identification problem and the general person recognition problem. In person recognition problem there is no constraint in the information that can be used. Motion and scene context information can also be exploited. Bounding box based re-identification approaches cannot extract these information. Face, Ear and Gait biometrics do not change significantly for a long period of time, whereas a person can change their clothing within a short period of time. It is not appropriate for person recognition applications to depend exclusively on the clothing information.

Movies capture large changes in viewpoint, illumination, and resolution in the images captured for any given characters. Also the clothing of the characters might change. Thus designing algorithms for recognition of characters on a movie will provide ample opportunity for solving real life re-identification problems. Another challenging problem is performing person recognition on family videos captured using hand held cameras. In these scenarios the person of interest might not be the focus of the camera. This creates new set of challenges for person recognition algorithms. University of Notre Dame has released a new dataset called “Point and Shoot Face Recognition dataset” [21] to solve these kinds of problems.
6. Conclusions
Most of the datasets used in re-identification are bounding box based, i.e., only a cropped image region of the person is available for both probe and gallery. This completely eliminates the use of potential context information. We believe that re-identification approaches should involve the use of the complete video information instead of the cropped image regions. Also there is no standard re-identification benchmark dataset currently available in the literature. The recently released video dataset as part of the ECCV workshops [19] is an attempt to create a standard benchmark dataset and the dataset suggests that the researchers are moving from bounding box based approach to video based approach. Currently, VIPeR is considered to be the difficult dataset and the rank-1 and rank-10 recognition accuracy on VIPeR is about 30% and 74% respectively [20]. Modelling the clothing information (current re-identification approaches) will improve the general person recognition. To solve the general person recognition problem researchers should focus on all the biometrics aspects like face, gait, ear, etc in the video data captured in the stationary as well as point and shoot cameras.

References
2. Hu, Yang, Dong Yi, Shengcai Liao, Zhen Lei, and Stan Z. Li. "Cross Dataset Person Re-identification."
15. http://gilscvblog.wordpress.com/2013/08/18/a-short-introduction-to-descriptors/
18. Javed, Omar; Shafique, Khurram; Shah, Mubarak. Appearance modeling for tracking in multiple non-overlapping cameras. CVPR, 2005