A framework for semantic people description in multi-camera surveillance systems

Zhou, Zhi; Wang, Yue; Teoh, Eam Khwang

2014

Zhou, Z., Wang, Y., & Teoh, E. K. (2015). A framework for semantic people description in multi-camera surveillance systems. Lecture Notes in Computer Science, 9010, 1-26.

<https://hdl.handle.net/10356/81702>

<https://doi.org/10.1007/978-3-319-16634-6>

© 2015 Springer International Publishing Switzerland. This is the author created version of a work that has been peer reviewed and accepted for publication by Computer Vision - ACCV 2014 Workshops, Lecture Notes in Computer Science, Springer. It incorporates referee's comments but changes resulting from the publishing process, such as copyediting, structural formatting, may not be reflected in this document. The published version is available at: [<http://dx.doi.org/10.1007/978-3-319-16634-6>].

Downloaded on 26 Apr 2023 23:23:36 SGT

Manuscript Number: IMAVIS-D-14-00285R2

Title: A framework for semantic people description in multi-camera
surveillance systems

Article Type: Full Length Article

Keywords: People re-identification; human appearance model; semantic
features; soft-biometric; surveillance.

Corresponding Author: Mr. Zhi Zhou,

Corresponding Author's Institution: Nanyang Technological University

First Author: Zhi Zhou

Order of Authors: Zhi Zhou; Yue Wang, Phd; Eam Khwang Teoh, Phd

Abstract: People re-identification has been a very active research topic recently in computer vision. It is an important application in surveillance systems with disjoint cameras. In this paper, a framework is proposed to extract descriptors of people in videos, which are based on soft-biometric traits and can be further used for people re-identification or other applications. Soft-biometric based description is more invariant to changing factors than directly using low level features such as color and texture. The ensemble of a set of soft-biometric traits can achieve good performance in people re-identification. In the proposed method, the body of detected people is divided into three parts and the selected soft-biometric traits are extracted from each part. All traits are then combined to form the final descriptor, and people re-identification is performed based on the descriptor and Nearest Neighbor (NN) matching strategy. The experiments are carried out on SAIVT-SoftBio database which consists of videos from disjoint surveillance cameras. An Open ID recognition problem is also evaluated for the proposed method. Comparisons with some state-of-the-art methods are provided as well. The experiment results show the good performance of the proposed framework.

Cover Letter

Dear Editor:

Please find enclosed a revised manuscript entitled “A framework for semantic people description in multi-camera surveillance systems” by Zhi Zhou, Yue Wang and Eam Khwang Teoh, Ref.: Ms. No. IMAVIS-D-14-00285R1, along with an attached reply to the reviewers’ report.

We appreciate very much the insightful comments made by the reviewers, and have revised the manuscript carefully according to all the reviewer’s comments. We believe that they have helped strengthen this paper and hope that this revised manuscript is now suitable for publication in Image and Vision Computing.

The minor revision as required is summarized as follows:

- 1) All comments from reviewers have been addressed and the modifications have been conducted in the revised version of our manuscript.
- 2) New experimental results for Open ID evaluation conducted on PRID dataset have been added in section 5.5.2.

A point-by-point response to the reviewer’s numbered comments is also included.

Thank you for your consideration.

Sincerely yours,

Zhi Zhou

People re-identification has been a very active research topic recently in computer vision. It is an important application in surveillance systems with disjoint cameras. In this paper, a framework is proposed to extract descriptors of people in videos, which are based on soft-biometric traits and can be further used for people re-identification or other applications. Soft-biometric based description is more invariant to changing factors than directly using low level features such as color and texture. The ensemble of a set of soft-biometric traits can achieve good performance in people re-identification. In the proposed method, the body of detected people is divided into three parts and the selected soft-biometric traits are extracted from each part. All traits are then combined to form the final descriptor, and people re-identification is performed based on the descriptor and Nearest Neighbor (NN) matching strategy. The experiments are carried out on SAIVT-SoftBio database which consists of videos from disjoint surveillance cameras. An Open ID recognition problem is also evaluated for the proposed method. Comparisons with some state-of-the-art methods are provided as well. The experiment results show the good performance of the proposed framework.

- No training or camera calibration is needed beforehand. As simple methods are developed for the extraction of soft-biometric traits, the proposed method is able to achieve real-time processing on videos.
- Able to handle the scale change and incomplete appearance. The proposed method can extract descriptors from images with different scales, even with some rotations. Descriptors of partially visible people can also be extracted, which is implausible in other state-of-the-art methods.
- No high resolution cameras are needed. Normal camera with resolution of 352x288 would meet the requirement.

A framework for semantic people description in multi-camera surveillance systems

Zhi Zhou^a, Yue Wang^b, Eam Khwang Teoh^a

^a*School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798*

^b*Visual Computing Department, Institute for Infocomm Research (I2R), Singapore 138632*

Abstract

People re-identification has been a very active research topic recently in computer vision. It is an important application in surveillance systems with disjoint cameras. In this paper, a framework is proposed to extract descriptors of people in videos, which are based on soft-biometric traits and can be further used for people re-identification or other applications. Soft-biometric based description is more invariant to changing factors than directly using low level features such as color and texture. The ensemble of a set of soft-biometric traits can achieve good performance in people re-identification. In the proposed method, the body of detected people is divided into three parts and the selected soft-biometric traits are extracted from each part. All traits are then combined to form the final descriptor, and people re-identification is performed based on the descriptor and Nearest Neighbor (NN) matching strategy. The experiments are carried out on SAIVT-SoftBio database which consists of videos from disjoint surveillance cameras, as well as some static image based datasets. An Open ID recognition problem is also evaluated for the proposed method. Comparisons with some state-of-the-art methods are provided as well. The experiment results show the good performance of the proposed framework.

Keywords: People re-identification, Human appearance model, Semantic features, Soft-biometric, Surveillance

1. Introduction

1.1. Background and motivation

Nowadays, camera based surveillance systems are widely used in our daily life. These surveillance systems are applied in many areas including home security, environment monitoring, industrial process control, surveillance of public places and so on. For some of these applications, the concern is the safety in public or private places with human as the main target. The traditional way is to hire human operators monitoring the system and report if any unexpected events or suspicious people are observed. However, as usual there are 20 to 40 cameras in one surveillance system, it is unrealistic and inefficiency for operators to monitor the whole system. To spare the human resource and reduce the cost, a surveillance system which is able to automatically detect the semantic information is highly desired. Tasks for such a system include detecting people in the video, tracking them and detecting their behaviors, even identifying the people if it is necessary.

To build such an automatic surveillance system, people detection and tracking algorithms are very important. Tracking people in one camera is plausible in practice. There are already some mature methods. For example, people

^{*}Corresponding author

Email addresses: zzhou5@e.ntu.edu.sg (Zhi Zhou), ywang@i2r.a-star.edu.sg (Yue Wang), eeekteoh@ntu.edu.sg (Eam Khwang Teoh)

can be tracked by using model matching [1], or searching in the neighborhood from last frame [2], or using joint probabilistic methods [3, 4]. However, surveillance systems in reality consist of several cameras covering different areas without overlapping. Compared with single camera based system, people tracking in multi-cameras based system is a more challenging work.

Consistently tracking one people in different cameras raises the demand to re-identify the people when he/she appears in different cameras or reappears in the same camera, known as people re-identification. People re-identification has been a hot-topic recently [5, 6]. However, it faces many problems. Firstly, due to the distribution of different cameras, the same person could appear in different view angles. This increases the difficulty of constructing a robust human appearance model. Secondly, there could be illumination variation in different cameras. Illumination variation is a challenging problem in people tracking and identification since it may affect the human appearance in the images. In addition, different cameras having different specifications could lead to a series of problems. For example, the same object may appear in slightly different colors in different cameras due to different image sensors used. Furthermore, the pose of people may keep changing while they are in some actions like walking or running. Another problem is the low resolution of images obtained in surveillance system. Last but not least, occlusion is a common problem existing in people tracking and identification.

Biometric information such as face [7] and gait [8] can do a great job in identifying people, but there are some constraints to use them in surveillance systems. To identify a people by face, it requires videos or images with high resolution and being captured in close distance. Also, the identification of face is mainly focusing on the frontal face, thus the cooperation of people is necessary. These are not suitable for the surveillance systems mentioned above. Moreover, different view angles of cameras and different poses of the same people make re-identification difficult by gaits.

Actually, the essential problem in people tracking or re-identification is to construct a model that can represent the people distinctly and robustly, especially under the condition that most of the camera based surveillance systems are covering different areas without overlap. Distinct means the discrepancy between the model of one people and models of other people should be as significant as possible, while robust means the model should be invariant to those changing factors. Besides using biometric information, one alternative is to model people by their appearance, assuming that the appearance of people does not change in short-term, such as hair color, skin tone and so on.

In this paper, a framework is proposed to construct human appearance model for people re-identification in disjointed camera surveillance system. Most existing people re-identification methods use local features like color, texture, key-points, etc., but there is a trend to include soft-biometric traits [9] in people re-identification. To form a distinctive yet robust descriptor, a set of soft-biometric traits are selected in the proposed framework. These traits describe certain parts of a people that do not change easily in short-term. Descriptors are further used for people re-identification. Though soft-biometric traits are not as distinct as biometric information in identifying a person, a set of them combined together are expected to achieve a good performance. Besides, an open ID problem is evaluated for the proposed method and some state-of-the-art methods, which have not been aware of in most existing methods.

The novelties of our proposed method are listed as below:

1. No training or camera calibration is needed beforehand. As simple methods are developed for the extraction of soft-biometric traits, the proposed method is able to achieve real-time processing on videos.
2. Able to handle the scale change and incomplete appearance. Unlike other methods extracting descriptors from well cropped and resized images, the proposed method can extract descriptors from images with different scales, even with some rotations. Descriptors of partially visible people can also be extracted, which is implausible in other state-of-the-art methods.
3. No high resolution cameras are needed. Normal camera with resolution of 352x288 would meet the requirement.

1.2. Paper organization

The rest of the paper is organized as follows. Section 2 gives a review on related works on people re-identification and soft-biometric. Overview of the proposed framework is given in section 3, followed by the detailed description of

the proposed method in section 4. Experimental results and discussion are provided in section 5. Finally, conclusion is addressed in section 6.

2. Related work

2.1. People re-identification methods based on human appearance

Recently, people re-identification has become an active topic in computer vision. People re-identification is known as the problem that after a person disappeared from one camera, he/she can be identified as the same person when he/she reappears in any camera of the multi-camera system. This is a common problem in a surveillance system with disjoint located cameras. In the past decade, some works have been done on people re-identification.

The research works on people re-identification can be roughly divided into two groups with different focuses.

(1) Focusing on the features used to construct the human model. Commonly used appearance features include color, texture, shape and so on.

Color is the most commonly used feature in many computer vision applications. One of the earliest works on people re-identification is from Javed [10]. The color histogram is used to represent a person. Color information is then feed into a Bayesian model together with the position and time information to track a person in disjoint cameras. Bird [11] segmented the human body into ten horizontal regions. For each region, the median values in Hue, Saturation and Lightness color spaces are calculated. The descriptor is formed by cascading value of median colors from each region. Then Fisher Linear Discriminants was used to perform appearance-based on-line classification. Kao etc. [12] used color information to form a hierarchical color structures for each image. For different people, a set of images is used and makes each node in the hierarchical color structures a Gaussian mixture model. Further calculation of the similarity of two people is based on Bayesian decision. In Madden's work [13], on-line clustering method is used to classify pixels belonging to a person. Then the color spectrum histogram is used to represent the person. Intensity transformation is applied when matching targets from disjoint cameras. Prosser [14] proposed Cumulative Brightness Transfer Function to make the color representation invariant to illumination changes due to different sites of disjoint cameras.

Texture is another popular feature used in people re-identification. Gray [15] picked an Ensemble of Localized Features (ELF). Features used in this method include texture features like Schmid and Gabor, as well as eight color channels from RGB, YCbCr and HSV color spaces. Farenzena [16] partitioned the person based on the image inside the silhouette. A person is segmented into head, torso and legs. Central axis is also calculated in torso and legs. Then for each part, local features concerning color and texture are used to represent this person. Features used in this method include weighted color histogram, Maximally Stable Color Regions (MSCR) and Recurrent High-Structured Patches (RHSP). Re-identification is done by matching those features in each part of the body. Berdugo [17] used background subtraction and saliency map to segment the people first, and employed three texture features—oriented gradients, color ratio and color saliency to form a discriminative descriptor. Ma [18] introduced Local Descriptors encoded by Fisher Vectors for people re-identification. A 7-d descriptor is calculated for each pixel including information on coordinate, intensity, first and second order derivatives. These descriptors are extracted on each channel of HSV color space. Then they are fit into Gaussian Mixture Model to obtain Fisher Vectors and concatenated to form the signature. Kviatkovsky [19] exploited a structure of color distributions by using different parts as well as shape context of the object in order to deal with illumination changes.

Besides color and texture, other features like local key-points, shape information and group information are also used in people re-identification. Gheissari [20] proposed a method combining interest point matching and model fitting in corresponding body parts. The body is segmented beforehand into different parts as head, torso and legs. Wang [21] introduced local shape information and appearance context for the representation of people. Histogram of Gradient (HOG) in the log-RGB color space is used as a local descriptor. Spatial distribution of the appearance inside each local part is considered as a feature as well. Hamdoun [22] extracted interest points from a person in a short period of time and use descriptors of interest points to model this person. Descriptors of interest points from different persons are used to build a K-D tree. Then for an unknown person, the identification is achieved by matching interest points in the K-D tree. Bak [23] used HOG to detect the human body and performed a color based foreground/background segmentation method. Two kinds of local features a set of haar-like features and Dominant Color Descriptor (DCD) are used. Mean Riemannian Covariance (MRC) patches are extracted from people and the combination of MRC patches

is used for people re-identification in [24]. Zheng [25] made use of information from close neighborhood to identify a person in group. Center Rectangular Ring Ratio-Occurrence is used to describe the global relation-ship of the group of people, and Block Based Ratio-Occurrence is used to describe local spatial information.

Methods mentioned above suffer a lot from changing factors like illumination and different view angles. There are also some soft-biometric traits used in people re-identification. These methods are reviewed in detail in section 2.2.

(2) Focusing on developing novel similarity calculation methods, or exploiting machine learning methods to enhance the people re-identification performance.

Gray [15] selected a set of local features and used AdaBoost to train samples. Optimal weights are obtained and assigned to each feature for the combination of final descriptor. Similarly, Prosser [26] used SVM to train samples with the same features used in [15]. In this method, people re-identification is considered as a ranking problem and a novel SVM-based ranking method is developed. In Bak's [23] method, AdaBoost is used to choose the most distinctive features from the feature set to represent the person. Avraham [27] used SVM to train the samples and obtain the transformation function between two cameras. Then for images in one camera, the descriptors can be matched with images in another camera after using the trained transformation function. Brand [28] extended the idea and made use of it on the people re-identification for more than two cameras. Wang [29] trained a distance metric with feature vectors extracted from overlapping areas. It showed good performance on people re-identification under occlusion and scale change. Zheng [30] also trained a metric by learning from samples. The probability is maximized that distances between true matches are smaller than false matches. Zhao [31] proposed a method by learning those salient features and match saliency distribution in datasets. Adjacency constrained patch matching is performed between two images and One-class SVM is used to find out salient score in the training set. Li [32] employed neural network to pick optimal features from a set of low level features for the re-identification task. A large dataset is built for testing of the method. Pedagadi [33] performs unsupervised Principle Component Analysis (PCA) and supervised Local Fisher Discriminant Analysis (LFDA) to extract the descriptor from the high dimensional features of color image. LDFA is employed to learn a distance metric between image pairs in the training set. In Tao's work [34], regularized smoothing KISS metric learning is proposed to extend Kostinger's work [35] to receive more stable result on small size training set.

Methods in the second group can deal with problems like illumination change. However, they need offline training for a specific scenario beforehand. In addition, large sample data is needed for better performance.

2.2. *Soft-biometric used in people re-identification*

In the recent years, soft-biometric has attracted more and more attention. Jain et al. [36, 37] defined soft-biometrics as those characteristics from human body that can provide robust but not identical information about the person. Dantcheva et al. [9] redefined soft-biometrics as physical, behavioral or adhered human characteristics. These traits are natural with semantic meaning, and can be used to describe humans. In this way, the range of soft-biometric is extended and not restricted to traits belonging to the human body. These soft-biometric include permanent traits like gender, ethnicity, or traits of human body appearance that do not change easily, such as age, skin color, hair color, eye color, body shapes and so on. Some appearance traits remain unchanged in a short-term time, though they are not biometrical part of human, are also considered as soft-biometric information. These traits include wearing of caps and glasses, or information from clothes and accessories.

Some of soft-biometric traits have already been used in people re-identification. In [38], more than twenty of soft-biometric traits like age, gender, race, skin color, hair color and so on were manually extracted by annotators from video. Another biometric feature used is the symmetry gait signatures. It describes the symmetry of the body in a period. Fourier decomposition is used to remove noise. Then Analysis of Variation (ANOVA) is used to select those most important traits to represent people. Shakhnarovich [39] proposed a framework to do some analysis after the face detected. Gender and ethnicity are two traits chosen. Some learning process based on AdaBoost or SVM are provided to train binary classifiers. Simple features such as Haar-like features are used. Denman [40] constructed a soft-biometric traits based model with height and color information in three body parts. They show acceptable performance on people re-identification in multi-camera scenarios. Vaquero [41] introduced a people identification method by using soft-biometric traits in the head area such as facial hair type, type of eyewear and hair type. Adaboost is used to train selected traits. Dantcheva and Dugelay [42] proposed a face re-identification method based on three soft-biometric traits hair, skin and clothes. For each trait, patches are retrieved from fixed areas. Color and texture

Figure 1. Overview of the proposed framework

are used to describe it. In Layne's works [43, 44], the human is segmented into 7 horizontal regions and 15 binary attributes were chosen by the author. All attributes are trained by SVM based on color and texture features in a specific region.

3. Our proposed framework for the construction of human appearance model– an overview

Currently, most of the existing methods on people re-identification are tested on static image datasets. These methods require images to be well aligned beforehand. However, in practice, real-time processing the video for people re-identification is expected. That demands methods can directly process on videos, which means images of people are not well segmented and they are in different scales. Also, people in camera may be tilted. Furthermore, learning beforehand for any two cameras may not be available. Therefore, in this paper, a framework that can automatically detect people in surveillance cameras and model the people with soft-biometric based traits is introduced. Image alignment and learning procedure are not involved beforehand.

The overview of the proposed framework is shown in Figure 1. Basically it consists of 4 steps.

Pre-processing

Before processing to the extraction of soft-biometric traits on people detected, some pre-processing works need to be done. In the proposed framework, background subtraction [45] and head detection method [46] are employed to detect people in the video. After each people detected, an ID is assigned to link the same people in consecutive frames. Body crop and rotation correction are included in pre-processing as well.

Body Segmentation

For each people detected in a frame, the body will be segmented into head, torso and leg areas. Symmetry and color information are both used for segmentation, which will be detailed in section 4.

Soft-biometric

There are many soft-biometric traits can be used. However, for demonstration purpose, 2 to 4 soft-biometric traits in each body part, 8 in total are used in the simulation of the proposed framework to construct the human appearance model. In the head area, soft-biometric traits like hair length and skin tone are extracted. In the torso area, traits describing the clothes are considered, including clothes color, clothes pattern and the length of sleeves. The detection of bag is performed as well. In the leg area, pant color and pant length are extracted.

Combination

After soft-biometric traits extracted, they are combined with weight to form an overall descriptor for a person. Re-identification work is further done by comparing the similarity of two sets of descriptors and assign IDs to images in the probe set. During the calculation of similarities, only those traits extracted are considered.

4. A novel approach of constructing human appearance model for people re-identification

In this section, the proposed framework is presented in detail. First, some pre-processing works are addressed, followed by segmentation method which divides the human body into three parts. Then methods extracting each soft-biometric trait are discussed. Finally, combination of soft-biometric traits and calculation of similarity in people re-identification are introduced.

4.1. Preprocessing works

Usually the video is transformed into frames for further processing. Considering the common frame sizes for surveillance cameras are 352×288 or 728×532 , the frames are resize to size 352×288 in the proposed framework if they are not.

4.1.1. People detection

Before constructing the human appearance model with soft-biometric traits, one important task is to detect people in the surveillance video. Existing methods in people detection include Histograms of Oriented Gradient (HOG) [47] and background subtraction based methods [45, 48]. HOG is a popular feature used for people detection in images. It describes the local shape information. Supervised learning methods like SVM are used to train HOG features to obtain a classifier. Background subtraction based methods are commonly used for mobile object detection in videos. A background model is constructed, and the foreground object in a frame with the same background can be detected by subtracting the background model.

HOG proves its accuracy in the detection of people. However, it has disadvantages. First, HOG is better to be used to decide whether human appears in a certain area. Usually bounding boxes are used for HOG based human detection, this means background information is inevitably included, which is not preferred in people re-identification because background information may distract the construction of human model. In addition, HOG requires a lot of calculation which is not suitable for real-time application. Considering these, background subtraction based method [45] is adopted along with head detection method [46] in this framework. Some of the background subtraction results are shown in Figure 2.

4.1.2. Body Crop

After the location of head is detected, a morphological operation is employed to crop the foreground mask of the human body. With the foreground mask given by background subtraction and the location given by head detection, the foreground mask area connected with head is cropped and considered as the mask of full body of people detected, as shown in Figure 2. Further processing like segmentation and extraction of soft-biometric traits are based on the cropped images.

4.1.3. Rotation correction

Considering images used come from surveillance systems in practice, affine transformation may occur sometimes (Figure 2). As the people appeared in image are not always in upright position, a rotation correction will be conducted if needed. The principle axis is first calculated based on the mask cropped. Principle Component Analysis (PCA) [49] is used to get the orientation of the principle axis. The first component is calculated as

$$v_1 = \operatorname{argmax}\left\{\frac{v^T X^T X v}{v^T v}\right\} \quad (1)$$

where $X = [x^T \ y^T]$ is the matrix storing location information for foreground pixels. v represents an eigenvector. Then the orientation can be computed with the first component $v_1 = [v_x \ v_y]$

$$\theta = \arctan\left(\frac{v_y}{v_x}\right) \quad (2)$$

Figure 2. Examples of the preprocessing work. Original images are shown in row 1. Row 2 are results of background subtraction, the target is shown in red bounding box. Row 3 are masks of cropped bodies. Rows 4 and 5 are masks and images of body after rotation correction

Then rotation transformation is applied to get the image of upright person.

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos(-\theta) & \sin(-\theta) \\ \sin(-\theta) & \cos(-\theta) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \quad (3)$$

4.2. Body segmentation

After the cropped image of human body obtained, the human body is to be divided into several areas for further extraction of soft-biometric traits. The conventional method is to segment the body into several horizontal stripes [11, 43], or segment the body into 3 parts: head, torso and leg [16, 41]. In the proposed framework, the latter segmentation is followed. A spatial distribution and color histogram combined methods which are similar to [16] are used for the segmentation.

The segmentation is based on the foreground mask which obtained from pre-processing. Segmentation between head and torso relies on the spatial distribution of mask area. For i_{th} row, the width of foreground area is denoted as w_i . The spatial distribution of two areas segmented by row i is calculated as

$$sd_i = var(W_h) + var(W_l) \quad (4)$$

where $W_h = [w_1, \dots, w_i]$ and $W_l = [w_{i+1}, \dots, w_n]$ are the sets of foreground width from the starting rows to the ending rows in two areas above and below the row i . Then the row which receives the minimum spatial distribution is searched. This row is taken as the border of head area and torso area.

$$r_{ht} = \underset{i}{\operatorname{argmin}}(sd_i) \quad (5)$$

where the search is constrained to the upper area of mask image.

For the segmentation between torso and leg, color information is used. Usually, the clothes color is not the same with the color of pants. Therefore, the distance between color histograms of upper area and lower area of human body is checked. For row i , the distance is calculated as

$$c_i = \|H_h - H_l\| \quad (6)$$

Figure 3. Examples of body segmentation. First row are human body masks, while second row are upper and lower body segmentation results.

where $\|\cdot\|$ represents Euclidean distance. H_h and H_l represent the color histograms in two areas, respectively. The row with maximum distance is considered as the border of torso and leg areas.

$$r_{tl} = \underset{i}{\operatorname{argmax}}(c_i) \quad (7)$$

To speed up the calculation, integral histogram is used when color histogram is formed. Some results of segmentation are shown in Figure 3.

Unlike the segmentation method used in [16] which can only conduct the segmentation on the full body, the method used in the proposed framework can be used to segment human body partially appears. For example, if only the head and torso can be seen from the frame, they can also be segmented without the leg area.

4.3. Extraction of soft-biometric traits

As introduced earlier, soft-biometric traits are those traits temporally adhere to human characteristic. They are with semantic meanings created by human to distinguish different people. Normally, one soft-biometric trait is not discriminative, but a set of them are proved to have good performance on people re-identification [43, 44]. In the proposed framework, some robust soft-biometric traits are chosen in simulation. The detection methods of these soft-biometric traits are simple yet fast, details are given as follows.

4.3.1. Head area

In the head area, there are many traits can be involved, such as hair, skin tone, wearing glasses and hat, etc. In the proposed framework, due to low video resolution used, only hair length and skin tone are chosen as the traits to describe the head area. Because they occupy larger areas than glasses and facial hairs and are detectable even in low resolution images. Of course, if high resolution video is available, more traits can be included.

Two categories are defined for the detection of hair length short hair and long hair. Hair length is extracted in the head patch, using areas with foreground mask. Blob analysis based method [50] is used to detect the hair length. Hair which is longer than the shoulder is considered as long hair. Otherwise it is classified as short hair. For detail please refer to our previous work in [50]. Skin tone also has two categories bright skin and dark skin. In the head patch, areas with foreground mask and excluded from hair area are considered as face area. Color histogram is extracted from face area and classified into bright skin and dark skin. Noted that the extraction of skin tone is highly depend on the segmentation of face area, therefore the detection rate is relatively low. Future work may exploit other segmentation method and other color spaces for better performance. Some examples of the traits in head area are shown in Figure 4.

4.3.2. Torso area

More traits are extracted from the torso area because usually this area provides more information than the other two areas. Four traits are used in the torso area in the proposed framework clothes color, clothes pattern, sleeve length and bag carrying.

Color is the most common used feature in computer vision due to its good ability to differentiate. However, images in disjointed surveillance cameras suffer from two aspects: (1) different illumination conditions; (2) cameras

Figure 4. Examples of soft-biometric traits in head area.

Figure 5. Examples of traits extraction in torso and leg areas. For the torso area, first row of texts denotes the clothes color, second row denotes sleeve length, third row denotes clothes pattern, the last row denotes bag carrying if exist. For the leg area, first row of texts denotes pant color, second row denotes pant length.

with different specifications or brands. The same color may appear differently in different cameras. To make the detection more robust, predefined color categories is used instead of color histograms in the conventional way. Colors of clothes are classified into nine common categories: black, gray, white, red, green, blue, cyan, magenta and yellow. Pixels in the torso area with foreground mask are considered belonging to the body. A color histogram is formed in HSV color space from these pixels. Color category with the maximum vote is considered as the dominant color and is used to represent the color of the clothes.

As often used by human when describing a people from distance, clothes pattern is used as a soft biometric trait in the proposed framework. There are many categories to describe the pattern of clothes, but it is not practical to use all the categories in the framework. Thus the most common three patterns: plain color, horizontal stripe and vertical stripe are chosen, the rest is categorized as complicate pattern. Therefore, there are four categories in total. Histograms of magnitude and orientation of gradient are extracted from pixels with foreground masks. Plain color is decided only by the histogram of magnitude of gradient. The rest categories are decided by both of the two histograms. For details please refer to our previous work in [51].

Another soft-biometric trait used is the length of sleeves. The length of sleeve is extracted indirectly by detecting nude arms. Just two categories are defined short sleeve and long sleeve. This is decided by whether large area of skin can be detected in the torso area. Skin can be detected by color since the skin color usually stays in a specific range [52].

Bag carrying can also be considered as a soft biometric trait. There are some works exist to detect the bag carrying in pictures [53, 54, 55]. In this framework, a simple method detecting the vertical Hough lines on the edge of the torso is used, since straps of bag usually have different color with clothes and form vertical lines.

Some examples of the traits in torso area are shown in Figure 5.

4.3.3. Leg area

In the leg area, pant color and pant length are extracted as soft biometric traits. For the extraction of pant color, similar method is used as in the extraction of clothes color. For the extraction of pant length, short pant and long pant are the defined categories, and similar method is used as in the extraction of sleeve length.

Some examples of the traits in leg area are shown in Figure 5.

4.4. Combination of soft-biometric traits and Re-identification

A descriptor for a person is a combination of soft-biometric traits extracted from three body parts

$$f = [v_1, \dots, v_t]^T \quad (8)$$

where v denotes a soft-biometric trait and t denotes the number of traits extracted. Most of existing datasets on people re-identification are consist of static images. That means the descriptor can only be extracted from a single image. However, it is highly possible that the image includes some unpredictable conditions which would downgrade the re-identification performance. The descriptor f_{is} extracted from single frame can be represented as

$$f_{is} = f_i + \epsilon \quad (9)$$

where ϵ represents unpredicted error and f_i is descriptor from the real model. Descriptor f_{im} extracted from a set of frames can be represented as

$$f_{im} = f_i + \frac{1}{n} \sum_{k=1}^n \epsilon_k \quad (10)$$

where n denotes the number of frames used to construct the descriptor. It can be an arbitrary number depends on different conditions. It is set to be 5 in the experiment part of the proposed method. Use σ to denote the variance of error. Suppose error ϵ follows normal distribution. The variance of error σ_{im} in multi-frame based descriptor is reduced compared with variance of error σ_{is} in single frame based descriptor

$$\sigma_{im} = \frac{1}{n} \sigma_{is} = \frac{1}{n} \sigma \quad (11)$$

Therefore, the descriptor constructed is based on the video instead of one static image. The video is first converted to a set of frames, then a set of descriptors are extracted based on frames. For a video, the descriptor for the people is obtained by weighted combination of all descriptors from frames. In multi-frame based descriptor, each trait is constructed as

$$v_k = \operatorname{argmax}(v_{kc}) \quad (12)$$

where k represents the k_{th} trait, and c is the number of categories of this trait exist in these frames.

Nearest Neighbor (NN) matching strategy is used in people re-identification experiments when assigning IDs to targets in the probe set. The similarity of two descriptors is calculated as

$$s(f_i, f_j) = \sum_{k=1}^t \alpha_k h(v_{ik}, v_{jk}) \quad (13)$$

$h(v_{ik}, v_{jk})$ equals to one if categories of the trait from two descriptors are the same, otherwise it equals to zero. The weight α_k is calculated as the entropy of trait category distributions in real-life. In the experiment, we use the trait category distributions in part of the dataset to calculate α_k for each trait. Then for a target in the probe set, the ID is determined by the person in the gallery set with closest distance:

$$ID_j = \operatorname{argmax}_{i \in 1, \dots, N} s(g_i, p_j) \quad (14)$$

5. Experimental results and discussion

To evaluate the performance of the proposed framework, experiments are carried out on a public dataset. There are several public datasets exist. Currently, the most commonly used datasets are VIPeR [56], iLIDS [57] and ETHZ [58], which consist of static image pairs for people. VIPeR and iLIDS contain static image pairs which are resized to 128×64 . Challenging conditions like illumination variations and occlusion are included. ETHZ dataset contains images of people cropped from videos taken by one moving camera. Multiple images of one person are taken from the same video sequence. Datasets mentioned above only contains static images for people. However, these static image based datasets have their disadvantages. Considering unpredictable conditions often occur in a single image, such as sudden illumination change, occlusion by background object, or occlusion caused by movement of limbs, these will distract the description from the real model. On the other hand, constructing the model with a set of frames in a video can reduce the error. Therefore, the experiment of the proposed method is carried out on SAIVT-SoftBio database

Figure 6. Scenarios of surveillance cameras in SAIVT-SoftBio database [59].

[59], a video based dataset recently published. This database consists of videos from eight cameras distributed inside a hotel. Scenarios of cameras are shown in Figure 6. For each person walking across the hotel, he or she may appear in one or several cameras. 152 people in total are included in this database.

To evaluate the re-identification performance, Cumulative Matching Characteristic (CMC) curve and Synthetic Recognition Rate (SRR) curve are employed. CMC curve records the probability that the right match can be found within the rank score. It is a cumulative curve that increases as the rank score increases. SRR curve records the probability that finds the right match at the first rank. This curve decreases as the number of targets increases. Area under the Curve (AUC) is provided for CMC curve as well. The proposed method is implemented with Matlab without optimization and tested on a computer with a dual core 2.8GHz, 16GB memory. It achieves 5 frames per second during processing.

There will be five parts in the experiment. Comparison with other methods on static image datasets is addressed in the first part. In the second part, the effects of multi-frames and single frame on the construction of model are investigated. In the third part, the performance of the proposed method on different viewing variations is evaluated, followed by the evaluation of the proposed method on all cameras. The last part discusses an open ID recognition problem.

5.1. Performance evaluation on static image datasets

First, we evaluate the performance of the proposed framework on some static image datasets and compare with other methods.

5.1.1. Comparison with other methods on VIPeR dataset

In this part, the proposed method is tested on VIPeR dataset and compared with SDALF [16], RDC [30] and SDC [31]. The codes for the comparing methods are provided by authors. For RDC and SDC, 316 pairs of images are randomly selected as training images, the rest are used as testing images. While SDALF and the proposed method randomly selected 316 pairs of images for the re-identification experiment. Figure 7 presents the CMC curve and SRR curve of the proposed method and other methods. As shown in the CMC curve, the proposed method achieves comparable performance with RDC, and outperforms SDALF and SDC. The proposed method also shows the best performance on the overall SRR curve.

AUCs are list in Table 1. RDC receives the best performance, while the proposed method receives a close score with RDC. Some statistics at lower rank scores of CMC curve are shown in Table 2. SDC receives best performance on rank 1, 5 and 10, while the proposed method achieves best performance on rank 25 and 50.

Table 1. Comparison of AUCs (in percentage) on VIPeR dataset, the best is shown in bold.

Method	SDALF	RDC	SDC	Proposed Method
AUCs	87.85	91.97	88.39	91.83

Figure 7. Comparison on CMC curve and SRR curve between the proposed method and other methods on VIPeR dataset

Table 2. Comparison of recognition rate (in percentage) in lower rank scores on VIPeR dataset, the best is shown in bold.

Method	SDALF	RDC	SDC	Proposed Method
Rank 1	13.54	17.32	24.05	20.68
Rank 5	32.82	35.69	42.72	38.64
Rank 10	45.95	45.39	50.63	49.15
Rank 25	66.87	62.68	69.62	70.17
Rank 50	84.05	76.78	79.75	86.78

5.1.2. Comparison with other methods on iLIDS dataset

In this part, the proposed method is tested on iLIDS dataset and compared with SDALF [16], RDC [30] and SDC [31]. For RDC and SDC, images from half of 119 subjects are randomly selected as training images, the rest are used as testing images. For SDALF and the proposed method, images from randomly selected 60 subjects are used in the re-identification experiment. Figure 8 presents the CMC curve and SRR curve of the proposed method and other methods. As shown in the both curves, the proposed method does not show superior performance on this dataset, while SDC receives the best performance.

AUCs are list in Table 3. SDC receives the best performance. The proposed method outperforms SDALF and RDC on iLIDS dataset. Some statistics at lower rank scores of CMC curve are shown in Table 4. Overall, SDC receives best performance.

Figure 8. Comparison on CMC curve and SRR curve between the proposed method and other methods on iLIDS dataset.

Table 3. Comparison of AUCs (in percentage) on iLIDS dataset, the best is shown in bold.

Method	SDALF	RDC	SDC	Proposed Method
AUCs	81.27	83.07	87.28	85.83

Table 4. Comparison of recognition rate (in percentage) in lower rank scores on iLIDS dataset, the best is shown in bold.

Method	SDALF	RDC	SDC	Proposed Method
Rank 1	30.51	24.33	42.17	34.24
Rank 5	55.59	51.67	62.33	58.47
Rank 10	68.14	63.17	75.33	71.02
Rank 20	85.25	79.17	84.50	83.56

As shown on the experimental results of the two static image datasets, the proposed method achieves comparable performance with other methods. However, it does not show superior performance on these datasets. Methods with supervised learning show their advantages on images pairs with similar viewing angles like in iLIDS dataset.

5.2. Single frame vs Multi-frames for human appearance model construction

In this section, the performances of descriptors formed by using multi-frames and single frame are discussed. Videos from all cameras are used. For single frame based descriptors of each people, a random frame from a random camera which is available is chosen. For multi-frame based descriptors of each people, 5 frames are randomly selected from one video and a single descriptor is formed by combining descriptors from these 5 frames. The CMC curve is shown in Figure 9. Multi-frames based descriptors show superior performance than single frame based descriptors in people re-identification. The curves are obtained with 10 times of repeated experiments. AUC is shown in Table 5, it also proves that multi-frames based descriptor receives better performance. The SRR curve is shown in Figure 9. It is obtained from a 100 times repeated experiment. From the SRR curve we can observe, the performance of multi-frames based descriptors is slightly better than single frame based descriptors. Therefore, we can conclude that using multiple frames based descriptors can improve the people re-identification performance.

Figure 9. CMC curve and SRR curve of single frame based descriptor and multi-frame based descriptor.

5.3. Evaluation of the performance on different view angles

In this section, the performance of the proposed framework on different view angles is evaluated. Comparison with single feature and other state-of-the-art methods is provided as well. The SAIVT-SoftBio database [59] provides videos from 8 cameras distributed in the hotel, which perfectly reflect view angle conditions in reality. Three conditions are analyzed in this part: cameras with similar views, cameras with opposite views and cameras with dissimilar views. Table 6 shows the view differences between each two cameras. In all these cameras, camera 2 captured too few subjects which are insufficient to be evaluated. Camera 4 suffered severe illumination condition. Camera 5 also has some illumination problem.

Table 5. AUC of single frame based descriptor and multi-frame based descriptor.

Method	Single Frame	Multi-frame
AUCs	78.62	87.18

Table 6. View differences between each pair of cameras in SAIVT-SoftBio database, where s represents similar views, o represents opposite views and d represents dissimilar views.

Cam No.	1	2	3	4	5	6	7	8
1	-	d	o	d	d	d	d	o
2	d	-	d	d	o	o	o	d
3	o	d	-	d	d	d	d	s
4	d	d	d	-	d	d	d	d
5	d	o	d	d	-	s	s	d
6	d	o	d	d	s	-	s	d
7	d	o	d	d	s	s	-	d
8	o	d	s	d	d	d	d	-

For each condition, the performances of soft-biometric traits combined descriptor with descriptors using single soft-biometric trait are first compared. For descriptors using single soft-biometric trait, continuous value without quantization is employed. Skin tone and Bag carrying are not included in the comparison. Bag carrying is a binary trait and not able to be represented by continuous value. The successful extraction of skin tone is quite low compared with other traits. Methods used for comparison includes SDALF [16], RDC [30] and SDC [31]. For existing methods, we use codes provided by authors for evaluation.

5.3.1. Condition 1: Similar views

Firstly, the effect of similar viewing angles is evaluated. In cameras with similar viewing angles, people captured usually have close angles of view, but other variations like scale and illumination may exist. Among the 8 cameras, there are two sets of similar view cameras. One set contains camera 3 and 8, the other set contains camera 5, 6 and 7.

The performances of multiple soft-biometric traits based descriptor with single soft-biometric trait based descriptor are first compared. CMC and SRR curves are shown in Figure 10. AUCs are listed in Table 7. Experiments are carried out on videos from a pair of two cameras with similar views. As can be seen in Figure 10, multi-traits based descriptor achieves the best performance in CMC curve. In SRR curve, multi-traits based descriptor is performing better than most of single trait based descriptors. Only the descriptor of hair length achieves higher detection rate on the SRR curve, this shows the discriminative of hair length when targets are few, it will decrease drastically as the number of targets increase which leads to the poor performance in CMC curve. Overall, multi-traits based descriptor outperforms descriptors extracted from single soft-biometric trait. This supports the view that, though single soft-biometric trait is not discriminative, a set of these traits can do well in identification.

The comparison of the proposed method with existing methods is shown in Figure 11. The AUCs are listed in Table 8. Noted that, the number of targets used in SRR curve might be less than 25 for RDC and SDC, this is because training is needed for RDC and SDC. Samples are divided into two groups, and half of the samples are used for training. Thus samples being used for test are less than 25. The performance of experiment on camera 5 and 7 is inferior to experiment on camera 3 and 8. This is due to the poor illumination condition in camera 5, which deteriorates the re-identification performance. Overall, the result of the proposed method has superior performance over the other three methods.

5.3.2. Condition 2: Opposite views

There are two sets of opposite cameras: (1) camera 1 vs camera 3 and 8; (2) camera 2 vs camera 5, 6, 7. However, due to insufficient subjects collected from camera 2, only the set camera 1 vs camera 3 and 8 is used in this experiment. With opposite views, people re-identification may face more problems than in similar viewing angles. Because even

Figure 10. CMC curve and SRR curve of multiple soft-biometric traits based descriptor and single feature based descriptors with similar views

Table 7. AUCs of multiple soft-biometric traits based descriptor and single feature based descriptors with similar views.

Features	AUCs(cam 3 and 8)	AUCs(cam 5 and 7)
Combined	84.16	79.93
Hair Length	53.05	64.38
Pant color	62.67	56.23
Clothes color	65.60	62.39
Pant length	55.13	54.65
Sleeve length	60.10	61.24
Clothes Pattern	73.38	76.72

Table 8. AUCs of proposed method, SDALF, RDC and SDC with similar views.

Method	SDALF	RDC	SDC	Proposed Method
AUCs(cam 3 vs cam 8)	76.90	83.81	77.20	84.16
AUCs(cam 5 vs cam 7)	69.14	72.09	73.25	79.93

Figure 11. CMC curve and SRR curve of proposed method, SDALF, RDC and SDC with similar views.

for the same person, some of the traits appear in the front view and the back view could be totally different. For example, there is a people wearing a white T-shirt and carrying a black backpack. From the front view, the clothes color is extracted as white, but from the back view, black may be extracted instead as the clothes color.

Similar with the previous condition, performances of multi-traits based descriptor and descriptors of single trait are first compared. CMC and SRR curves are shown in Figure 12. AUCs are listed in Table 9. From the figure and table, it can be concluded that multi-traits based descriptor outperforms descriptors of single trait in people re-identification of opposite views.

Then the proposed method is compared with existing methods on opposite viewing angles. The comparison is shown in Figure 13 and the AUCs are listed in Table 10. It can be observed that the proposed method is performing better. Compared with the performance of all methods in similar views (camera 3 and 8), the AUCs of all methods declined in opposite views. This denotes that people re-identification is more challenging in opposite views than in

Figure 12. CMC curve and SRR curve of multiple soft-biometric traits based descriptor and single feature based descriptors with opposite views.

Table 9. AUCs of multiple soft-biometric traits based descriptor and single feature based descriptors with opposite views.

Features	AUCs(cam 1 vs cam 3 and 8)
Combined	78.96
Hair Length	57.80
Pant color	58.01
Clothes color	60.22
Pant length	59.73
Sleeve length	68.21
Clothes Pattern	66.65

Figure 13. CMC curve and SRR curve of proposed method, SDALF, RDC and SDC with opposite views.

similar views.

5.3.3. Condition 3: Dissimilar views

There are more combinations for cameras with dissimilar views. Considering the views of angle are different, the performance could be inferior to that of similar views.

As the same with previous conditions, multi-traits based descriptor is compared with single trait based descriptors. CMC and SRR curves are shown in Figure 14 and the AUCs are listed in Table 11.

Then the proposed method is compared with existing methods on dissimilar viewing angles. The comparison is shown in Figure 15 and the AUCs are listed in Table 12. Again, as can be observed, the proposed method is performing better.

5.3.4. Discussion

Comparing the performance in three different viewing conditions, it is evident that experiments on similar views have the best re-identification performance. In experiments on opposite views and dissimilar views, the performance of all methods declines. It can be concluded that the view of camera is very important in people re-identification. Large variation in views has great impact on the performance. Among the four methods, RDC declines the most from similar views to other viewing conditions. RDC and SDC are receiving inferior performance on opposite views and

Table 10. AUCs of proposed method, SDALF, RDC and SDC with opposite views.

Method	SDALF	RDC	SDC	Proposed Method
AUCs(cam 1 vs cam 3 and 8)	75.45	73.32	71.44	78.96

Figure 14. CMC curve and SRR curve of multiple soft-biometric traits based descriptor and single feature based descriptors with dissimilar views.

Table 11. AUCs of multiple soft-biometric traits based descriptor and single feature based descriptors with dissimilar views.

Features	AUCs(cam 4 vs cam 5,6,7)	AUCs(cam 5,6,7 vs cam 3,8)
Combined	70.85	79.10
Hair Length	53.17	55.73
Pant color	52.69	58.69
Clothes color	55.36	64.80
Pant length	51.07	57.64
Sleeve length	59.76	61.36
Clothes Pattern	59.39	71.77

Table 12. AUCs of proposed method, SDALF, RDC and SDC with dissimilar views.

Method	SDALF	RDC	SDC	Proposed Method
AUCs(cam 4 vs cam 5,6,7)	62.17	61.50	58.18	70.85
AUCs(cam 5,6,7 vs cam 3,8)	72.07	68.54	71.94	79.10

Figure 15. CMC curve and SRR curve of proposed method, SDALF, RDC and SDC with dissimilar views.

dissimilar views. This may indicate that learning based methods performance largely depends on the viewing angles, especially when the sample data are insufficient.

5.4. Comparison with other methods with multi-views

People re-identifications in previous section are based on two cameras. What if the images in the gallery and probe are both from different viewing angles? In this section, evaluation of the performance of the proposed method under this condition is provided and compared with some existing methods. Different with last section, images of gallery and probe sets are randomly chosen from several cameras, without considering the view angles. From Figure 16 and Table 13, it is observed that with more view angles involved, the performance of the proposed method is still better than the other three methods.

Figure 16. CMC curve and SRR curve of proposed method, SDALF, RDC and SDC with multiple views.

Table 13. AUCs of proposed method, SDALF, RDC and SDC with multiple views.

Method	SDALF	RDC	SDC	Proposed Method
AUCs(all cameras)	73.45	64.03	69.48	78.59

Table 14. Optimum RDFPs obtained from proposed method, SDALF, RDC and SDC with similar views.

Method	SDALF	RDC	SDC	Proposed Method
RDFP(cam 3 vs cam 8)	50.00	100.00	9.79	33.33
RDFP(cam 5 vs cam 7)	50.00	20.00	14.51	22.22

5.5. Evaluation on Open ID recognition

Currently, most existing people re-identification experiments are based on gallery and probe sets containing exactly the same subjects. That means for each image in the probe set, there must be at least one image in the gallery set belonging to the same ID. This is defined as closed set experiment [60]. However, what faced in reality are open set problems [60], this means some people appear in the probe set are never included in the gallery set, vice versa.

The way to evaluate the performance in open set experiment is different with the closed set experiment. For the closed set experiment, the matching accuracy is the only index matters. However, for open set problems, besides matching accuracy, we have to care about the detection of outliers, which denotes the rate of mismatching for those images appear in only probe set.

Usually in the closed set experiment, the distance of two descriptors is measured to represent the similarity of two subjects. For one probe subject, the subject in gallery with the closest distance is picked as the matching ID. This method is not suitable for open set problem. Because in an open set problem, for a subject appears only in probe set, there is no right matching in the gallery set. A way is needed to judge whether the subject is an outlier or not. In open set experiment, a threshold could be set for the distance of two descriptors to decide outliers. The matching is confirmed only if the similarity of two subjects is larger than the threshold τ

$$s(g_i, p_j) > \tau \quad (15)$$

In this section, performances of the proposed method with existing methods SDALF [16], RDC [30] and SDC [31] are compared for open ID experiment on SAIVT-SoftBio database [59] and PRID dataset [61]. CMC curve is used, as well as Ratio of Detection rate to False Positive rate (RDFP) and Ratio of Detection rate to False Acceptance Rate (RDFAR). False positive denotes outliers from probe which were falsely matched as one of subjects in gallery. Mis-Matches denote non-outliers from probe which were assigned a false ID. False Acceptance Rate includes both False Positive rate and Mis-Match rate. Ratio of Detection rate to False Positive rate (RDFP) is defined as

$$RDFP = \frac{DetectionRate}{DetectionRate + FalsePositiveRate} \quad (16)$$

Ratio of Detection rate to False Acceptance Rate (RDFAR) is defined as

$$RDFAR = \frac{DetectionRate}{DetectionRate + FalsePositiveRate + MisMatchRate} \quad (17)$$

5.5.1. Open ID experiment on SAIVT-SoftBio database

Three scenarios—similar views, opposite views and dissimilar views are considered in SAIVT-SoftBio database. As can be seen from Figures 17–19, compared with closed ID experiment, the performances of all methods drop in open ID experiment. This is due to the distraction of unmatched samples. In open ID experiment the detection of outliers is more cared about, which is not reflected in CMC curve. Therefore, the RDFP and RDFAR are used. Since a fixed threshold is not desirable for different camera combinations, a varying threshold is used. In Figure 17–19, the threshold varies from the minimum value to the maximum value of similarity scores. 100 values are set for the varying threshold.

Figure 17. CMC curve, RDPF curve and RDFAR curve of proposed method, SDALF, RDC and SDC with similar views

Table 15. Optimum RDFARs obtained from proposed method, SDALF, RDC and SDC with similar views.

Method	SDALF	RDC	SDC	Proposed Method
RDFAR(cam 3 vs cam 8)	7.25	12.50	7.71	9.09
RDFAR(cam 5 vs cam 7)	6.18	6.06	7.22	12.80

Figure 18. CMC curve, RDPF curve and RDFAR curve of proposed method, SDALF, RDC and SDC with opposite views.

Table 16. Optimum RDFPs obtained from proposed method, SDALF, RDC and SDC with opposite views.

Method	SDALF	RDC	SDC	Proposed Method
RDFP(cam 1 vs cam 3 and 8)	19.05	57.14	20.53	37.50

Table 17. Optimum RDFARs obtained from proposed method, SDALF, RDC and SDC with opposite views.

Method	SDALF	RDC	SDC	Proposed Method
RDFAR(cam 1 vs cam 3 and 8)	4.88	15.15	6.81	8.26

Figure 19. CMC curve, RDFP curve and RDFAR curve of proposed method, SDALF, RDC and SDC with dissimilar views.

Table 18. Optimum RDFPs obtained from proposed method, SDALF, RDC and SDC with dissimilar views.

Method	SDALF	RDC	SDC	Proposed Method
RDFP(cam 1 vs cam 5,6,7)	16.47	50.00	14.00	45.45
RDFP(cam 4 vs cam 5,6,7)	2.88	14.29	3.65	14.29

Table 19. Optimum RDFARs obtained from proposed method, SDALF, RDC and SDC with dissimilar views.

Method	SDALF	RDC	SDC	Proposed Method
RDFAR(cam 4 vs cam 5,6,7)	3.98	9.26	2.17	8.13
RDFAR(cam 5,6,7 vs cam 3,8)	2.21	5.00	2.80	2.96

Table 20. Optimum RDFPs obtained from proposed method, SDALF, RDC and SDC on PRID dataset.

Method	SDALF	RDC	SDC	Proposed Method
RDFP(cam a vs cam b)	2.87	2.47	2.47	1.27
RDFP(cam b vs cam a)	1.19	7.13	0.78	4.69

As can be seen from Figures 17–19, RDC does not have good performance on CMC curve. However, its performance on RDFP and RDFAR is the best most of the time. This denotes RDC has good performance on the detection of outliers. RDC is a supervised learning based method. With learning progress, it gives low similarity score to outliers. Therefore the false positive which is not preferred in open ID experiment is kept low, though the detection rate of RDC is low. Meanwhile, SDC does not receive good performance on RDFP and RDFAR, the reason might be that SDC depends on finding the salient features of the target. However, in SAIVT-SoftBio database, appearances of most subjects are not salient to others. This cause the low detection rate for SDC even training is involved.

Another reason contributes to the superior performance of RDC in RDFP and RDFAR is that, subjects used in testing are less than subjects used in SDALF and the proposed method, due to the training data needed. With less subjects involved in experiment, the ratio of detection rate is supposed to be higher. Therefore, the proposed method is more proper to compare with SDALF, which is under the same experiment condition. As shown in Figures 17–19 and Table 14–19, the proposed method performs better than SDALF. Even though, the proposed method still faces the problem of excluding outliers when performing re-identification in an open ID experiment.

5.5.2. Opend ID experiment on PRID dataset

We also evaluated the performance of methods on PRID dataset [61], which is based on static images. There are two camera views in PRID dataset, one containing 385 subjects, while the other containing 749 subjects. 200 subjects among them appear in both camera views. Images in the two different camera views are in similar views.

As shown in Figure 20, the proposed method achieves comparable performance with SDC and outperforms the other two methods on the CMC curve. SDC receives best performance on RDFP and RDFAR overall, due to the similar views between two cameras. The proposed method also receives satisfying performance as the difference between two views is not significant. Optimum RDFPs and RDFARs obtained by methods are listed in table 20–21.

6. Conclusion

In this paper, a framework is proposed to extract soft-biometric based people descriptors from videos. This framework is designed to solve people re-identification problem in disjoint camera based surveillance systems. People detection, body segmentation and extraction of soft-biometric traits are included in the framework. The performance of the extracted descriptors is evaluated on several aspects and comparisons with state-of-the-art methods are provided. The experimental results show the good performances of the proposed method under conditions of different view angles, illumination change and pose variations. However, in open ID experiment, it faces the problem of excluding outliers, which should be considered and solved in future work. In future, more semantic features should be included into the framework to achieve more accurate description. Those soft-biometric traits requiring good resolution (wearing of glasses, facial hair, etc.) can be extracted when resolution condition is satisfied. Otherwise they are excluded from the descriptor. In addition, a multi-layer descriptor could be built. Semantic feature based descriptor can be used for rough re-identification. After that, local features could be used which may improve the performance of outlier detecting, thus a better performance can be expected in open ID experiment. Besides these, more datasets and existing methods will be involved in the experiments, to consolidate the evaluation.

Acknowledgement

We would like to thank the SAIVT Research Labs at Queensland University of Technology (QUT) for freely supplying us with the SAIVT-SoftBio database for our research.

Figure 20. CMC curve, RDFF curve and RDFAR curve of proposed method, SDALF, RDC and SDC on PRID dataset.

Table 21. Optimum RDFARs obtained from proposed method, SDALF, RDC and SDC on PRID dataset.

Method	SDALF	RDC	SDC	Proposed Method
RDFAR(cam a vs cam b)	0.58	0.76	1.67	0.60
RDFAR(cam b vs cam a)	0.21	0.76	0.32	0.38

References

- [1] M. Andriluka, S. Roth, B. Schiele, People-tracking-by-detection and people-detection-by-tracking, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–8.
- [2] S. Denman, V. Chandran, S. Sridharan, An adaptive optical flow technique for person tracking systems, Pattern Recognition Letters 28 (10) (2007) 1232–1239.
- [3] D. Schulz, W. Burgard, D. Fox, A. B. Cremers, People tracking with mobile robots using sample-based joint probabilistic data association filters, The International Journal of Robotics Research 22 (2) (2003) 99–116.
- [4] D. Schulz, D. Fox, J. Hightower, People tracking with anonymous and id-sensors using rao-blackwellised particle filters, in: IJCAI, 2003, pp. 921–928.
- [5] G. Doretto, T. Sebastian, P. Tu, J. Rittscher, Appearance-based person reidentification in camera networks: problem overview and current approaches, Journal of Ambient Intelligence and Humanized Computing 2 (2) (2011) 127–151.
- [6] A. Bedagkar-Gala, S. K. Shah, A survey of approaches and trends in person re-identification, Image and Vision Computing 32 (4) (2014) 270–286.
- [7] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, W. Worek, Overview of the face recognition grand challenge, in: CVPR, Vol. 1, 2005, pp. 947–954.
- [8] A. Bissacco, S. Soatto, Hybrid dynamical models of human motion for the recognition of human gaits, International journal of computer vision 85 (1) (2009) 101–114.
- [9] A. Dantcheva, C. Velardo, A. Dangelo, J.-L. Dugelay, Bag of soft biometrics for person identification, Multimedia Tools and Applications 51 (2) (2011) 739–777.
- [10] O. Javed, Z. Rasheed, K. Shafique, M. Shah, Tracking across multiple cameras with disjoint views, in: Proceedings of Ninth IEEE International Conference on Computer Vision (ICCV), 2003, pp. 952–957.
- [11] N. D. Bird, O. Masoud, N. P. Papanikolopoulos, A. Isaacs, Detection of loitering individuals in public transportation areas, IEEE Transactions on Intelligent Transportation Systems 6 (2) (2005) 167–177.
- [12] J.-H. Kao, C.-Y. Lin, W.-H. Wang, Y.-T. Wu, A unified hierarchical appearance model for people re-identification using multi-view vision sensors, in: Advances in Multimedia Information Processing-PCM, 2008, pp. 553–562.
- [13] C. Madden, E. D. Cheng, M. Piccardi, Tracking people across disjoint camera views by an illumination-tolerant appearance representation, Machine Vision and Applications 18 (3-4) (2007) 233–247.
- [14] B. Prosser, S. Gong, T. Xiang, Multi-camera matching using bi-directional cumulative brightness transfer functions, in: Proc. British Machine Vision Conference (BMVC), Vol. 8, 2008, pp. 164–1.

- [15] D. Gray, H. Tao, Viewpoint invariant pedestrian recognition with an ensemble of localized features, in: Proc. European Conference on Computer Vision (ECCV), 2008, pp. 262–275.
- [16] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person re-identification by symmetry-driven accumulation of local features, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 2360–2367.
- [17] G. Berdugo, O. Soceanu, Y. Moshe, D. Rudoy, I. Dvir, Object reidentification in real world scenarios across multiple non-overlapping cameras, in: Proc. Euro. Sig. Proc. Conf, 2010, pp. 1806–1810.
- [18] B. Ma, Y. Su, F. Jurie, Local descriptors encoded by fisher vectors for person re-identification, in: ECCV Workshops and Demonstrations, 2012, pp. 413–422.
- [19] I. Kviatkovsky, A. Adam, E. Rivlin, Color invariants for person reidentification, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (7) (2013) 1622–1634.
- [20] N. Gheissari, T. B. Sebastian, R. Hartley, Person reidentification using spatiotemporal appearance, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 2, 2006, pp. 1528–1535.
- [21] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, P. Tu, Shape and appearance context modeling, in: In: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2007.
- [22] O. Hamdoun, F. Moutarde, B. Stanciulescu, B. Steux, Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences, in: Second ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC), 2008, pp. 1–6.
- [23] S. Bak, E. Corvee, F. Brémont, M. Thonnat, Person re-identification using haar-based and dcd-based signature, in: IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2010, pp. 1–8.
- [24] S. Bak, E. Corvee, F. Brémont, M. Thonnat, Boosted human re-identification using riemannian manifolds, Image and Vision Computing 30 (6) (2012) 443–452.
- [25] W.-S. Zheng, S. Gong, T. Xiang, Associating groups of people, in: Proc. British Machine Vision Conference (BMVC), 2009.
- [26] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, Q. Mary, Person re-identification by support vector ranking., in: Proc. British Machine Vision Conference (BMVC), Vol. 1, 2010, p. 5.
- [27] T. Avraham, I. Gurvich, M. Lindenbaum, S. Markovitch, Learning implicit transfer for person re-identification, in: ECCV Workshops and Demonstrations, 2012, pp. 381–390.
- [28] Y. Brand, T. Avraham, M. Lindenbaum, Transitive re-identification, in: British Machine Vision Conference (BMVC), 2013.
- [29] S. Wang, M. Lewandowski, J. Annesley, J. Orwell, Re-identification of pedestrians with variable occlusion and scale, in: IEEE International Conference on Computer Vision Workshops (ICCV Workshops), 2011, pp. 1876–1882.
- [30] W.-S. Zheng, S. Gong, T. Xiang, Reidentification by relative distance comparison, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (3) (2013) 653–668.
- [31] R. Zhao, W. Ouyang, X. Wang, Unsupervised salience learning for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3586–3593.
- [32] W. Li, R. Zhao, T. Xiao, X. Wang, Deepreid: Deep filter pairing neural network for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 152–159.
- [33] S. Pedagadi, J. Orwell, S. Velastin, B. Boghossian, Local fisher discriminant analysis for pedestrian re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3318–3325.
- [34] D. Tao, L. Jin, Y. Wang, Y. Yuan, X. Li, Person re-identification by regularized smoothing kiss metric learning, IEEE Transactions on Circuits and Systems for Video Technology 23 (10) (2013) 1675–1685.
- [35] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, H. Bischof, Large scale metric learning from equivalence constraints, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 2288–2295.
- [36] A. K. Jain, S. C. Dass, K. Nandakumar, Soft biometric traits for personal recognition systems, in: Biometric Authentication, 2004, pp. 731–738.
- [37] A. K. Jain, S. C. Dass, K. Nandakumar, Can soft biometric traits assist user recognition?, in: Defense and Security, International Society for Optics and Photonics, 2004, pp. 561–572.
- [38] S. Samangoei, B. Guo, M. S. Nixon, The use of semantic human description as a soft biometric, in: 2nd IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), 2008, pp. 1–7.
- [39] G. Shakhnarovich, P. A. Viola, B. Moghaddam, A unified learning framework for real time face detection and classification, in: Proceedings. Fifth IEEE International Conference on Automatic Face and Gesture Recognition, 2002, pp. 14–21.
- [40] S. Denman, C. Fookes, A. Bialkowski, S. Sridharan, Soft-biometrics: unconstrained authentication in a surveillance environment, in: Digital Image Computing: Techniques and Applications (DICTA), 2009, pp. 196–203.
- [41] D. A. Vaquero, R. S. Feris, D. Tran, L. Brown, A. Hampapur, M. Turk, Attribute-based people search in surveillance environments, in: Workshop on Applications of Computer Vision (WACV), 2009, pp. 1–8.
- [42] A. Dantcheva, J.-L. Dugelay, Frontal-to-side face re-identification based on hair, skin and clothes patches, in: IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS), 2011, pp. 309–313.
- [43] R. Layne, T. M. Hospedales, S. Gong, et al., Person re-identification by attributes., in: Proc. British Machine Vision Conference (BMVC), Vol. 2, 2012, p. 3.
- [44] R. Layne, T. M. Hospedales, S. Gong, Towards person identification and re-identification with attributes, in: ECCV Workshops and Demonstrations, 2012, pp. 402–412.
- [45] L. Li, W. Huang, I. Y. Gu, Q. Tian, Foreground object detection from videos containing complex background, in: Proceedings of the eleventh ACM international conference on Multimedia, ACM, 2003, pp. 2–10.
- [46] S. Liao, X. Zhu, Z. Lei, L. Zhang, S. Z. Li, Learning multi-scale block local binary patterns for face recognition, in: Advances in Biometrics, 2007, pp. 828–837.
- [47] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 1, 2005, pp. 886–893.

- [48] H. Wang, D. Suter, A consensus-based method for tracking: Modelling background scenario and foreground appearance, *Pattern Recognition* 40 (3) (2007) 1091–1105.
- [49] I. Jolliffe, *Principal component analysis*, Wiley Online Library, 2002.
- [50] Y. Wang, Z. Zhou, E. K. Teoh, Human hair segmentation and length detection for human appearance model, in: *International Conference on Pattern Recognition (ICPR)*, 2014.
- [51] Z. Zhou, Y. Wang, E. K. Teoh, People apparel model for human appearance description, in: *International Conference on Information, Communications and Signal Processing (ICICSP)*, 2013, pp. 1–5.
- [52] K. Sobottka, I. Pitas, A novel method for automatic face segmentation, facial feature extraction and tracking, *Signal processing: Image communication* 12 (3) (1998) 263–281.
- [53] I. Haritaoglu, R. Cutler, D. Harwood, L. S. Davis, Backpack: Detection of people carrying objects using silhouettes, in: *The Proceedings of the Seventh IEEE International Conference on Computer Vision*, Vol. 1, 1999, pp. 102–107.
- [54] D. Damen, D. Hogg, Detecting carried objects from sequences of walking pedestrians, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34 (6) (2012) 1056–1067.
- [55] C.-S. Lee, A. Elgammal, Carrying object detection using pose preserving dynamic shape models, in: *Articulated Motion and Deformable Objects*, 2006, pp. 315–325.
- [56] D. Gray, S. Brennan, H. Tao, Evaluating appearance models for recognition, reacquisition, and tracking, in: *IEEE International workshop on performance evaluation of tracking and surveillance*, 2007.
- [57] U. H. Office, i-LIDS multiple camera tracking scenario definition, 2008.
- [58] A. Ess, B. Leibe, L. Van Gool, Depth and appearance for mobile scene analysis, in: *IEEE 11th International Conference on Computer Vision (ICCV)*, 2007, pp. 1–8.
- [59] A. Bialkowski, S. Denman, P. Lucey, S. Sridharan, C. B. Fookes, A database for person re-identification in multi-camera surveillance networks, in: *Proceedings of the 2012 International Conference on Digital Image Computing Techniques and Applications (DICTA)*, 2012, pp. 1–8.
- [60] A. Bedagkar-Gala, S. K. Shah, Part-based spatio-temporal model for multi-person re-identification, *Pattern Recognition Letters* 33 (14) (2012) 1908–1915.
- [61] M. Hirzer, C. Belezni, P. M. Roth, H. Bischof, Person re-identification by descriptive and discriminative classification, in: *Image Analysis*, Springer, 2011, pp. 91–102.

Response to Comments

Image and Vision Computing

Ref.: Ms. No. IMAVIS-D-14-00285R1

Title A framework for semantic people description in multi-camera surveillance systems

Reviewer ID: Reviewer #1

Reviewer's Comment:

The authors have answered my previous concerns. But I have one more concern on section 5.5, which proposed openID evaluation, and used a thresholding approach to evaluate all methods. Actually there is a more suitable dataset named PRID [a] for the openID evaluation. The PRID dataset contains unbalanced number of persons in gallery and probe set, including the case that some people in probe set are not appeared in gallery set. Therefore, I would expect authors to add further experiments to evaluate the proposed method on the PRID dataset.

[a] Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof. Person Re-Identification by Descriptive and Discriminative Classification. In Proc. Scandinavian Conference on Image Analysis (SCIA), 2011.

Authors' Response:

Reviewer's suggestion is much appreciated, PRID dataset is suitable for OpenID experiment. Therefore, we have added experiments on PRID dataset to evaluate the proposed methods as well as methods for comparison. The added experiment is put in subsection 5.5.2 in the revised manuscript.

LaTeX Source Files

[Click here to download LaTeX Source Files: IMAVIS-D-14-00285R1_revised.zip](#)