Appearance-Invariant Entry-Exit Matching Using Visual Soft Biometric Traits

V. Vinay Kumar[1(✉)] and P. Nagabhushan[2]

[1] Department of Studies in Computer Science, University of Mysore, Mysuru, India
[2] Indian Institute of Information Technology Allahabad, Prayagraj, India

Abstract. The problem of appearance invariant subject re-identification for Entry-Exit surveillance applications is addressed. A novel Semantic Entry-Exit matching model that makes use of ancillary information about subjects such as height, build, complexion, and clothing color to endorse exit of every subject who had entered private area is proposed in this paper. The proposed method is robust to variations in appearances such as clothing, carrying, and head masking. Each describing attribute is given equal weight while computing the matching score, and hence the proposed model achieves high rank-k accuracy on benchmark datasets. The soft biometric traits used as a combination though, cannot achieve high rank-1 accuracy, it helps to narrow down the search to match using reliable biometric traits such as gait and face whose learning and matching time is costlier when compared to the soft biometrics.

Keywords: Entry-exit surveillance · Appearance-invariant person re-identification · Camera forbidden zones

1 Introduction

Intelligent video surveillance systems overcame the limitations in the human ability to diligently watch and monitor multiple live video surveillance footages [1]. The intelligent video surveillance domain witnessed extensive research in the past two decades, thus extending its applications to intruder detection and validation, crime prevention, elderly people, and children monitoring. Today, public places such as shopping malls, airports, buses, and rail stations are completely under surveillance ambit except in few areas such as toilets and changing rooms, which are referred to as *private areas* where installing surveillance cameras is considered a breach of privacy. This is often seen as a hindrance to security systems in crime prevention and public safety. As a solution, the notion of Entry-Exit surveillance (EES) [2] deals with the monitoring of subjects entering and exiting private areas. The key objective is to assure that the subjects who had entered private areas exit in time without much variations in their appearances that may lead to suspicion. Every subject who enters the private area is labeled and saved in the gallery, and every subject who exits the private area is considered as a probe and has to be matched with subjects available in the gallery.

© Springer Nature Singapore Pte Ltd. 2021
S. K. Singh et al. (Eds.): CVIP 2020, CCIS 1376, pp. 260–269, 2021.
https://doi.org/10.1007/978-981-16-1086-8_23

1.1 Relevance of the Problem

The problem of entry-exit matching can be related to a person re-identification problem where the aim is to match subjects moving between surveillance ambits of non-overlapping cameras. For every probe subject, the matching subject in the gallery set should have a high matching rank when compared with other subjects. A detailed survey on person re-identification can be found in [3,4]. In conventional person re-identification systems, it is assumed that the subject is to be matched across camera views on the same day, and hence the issue of variation in the appearance of subjects due to change in clothing is under-addressed. However, in entry-exit surveillance, the temporal gap between entry and exit where the subjects move out of surveillance ambit envisage the possibility of change in the appearance of subjects with respect to clothing, carrying, and head masking. Traces of appearance invariant subject re-identification solutions available in the literature include analysis of gait and motion patterns in [5]. However, state of the art gait recognition algorithms suffers due to variations in the walking directions. Face biometric trait is another promising attribute for subject re-identification that can be robust to changes in clothing and carrying conditions. However, capturing of face attributes is limited to very few frames due to the distance of subjects from the camera. Adding to it, the possibility of face occlusions due to subjects overlapping as well as religious and cultural practices in unconstrained environments limits the extraction of face attributes from surveillance videos.

On the other hand, visual soft biometric attributes such as clothing color, height, body-build, accessories possessed by the subjects can be extracted from low-resolution video frames. Most of the current state of the art person re-identification methods mainly focus on clothing attributes for matching. However, in entry-exit surveillance, due to the possibility of variations in the clothing of subjects, relatively less importance must be attributed to clothing color but cannot be completely dropped as it has a high discriminating ability in the majority of the cases. The probability of change in clothing in places such as toilets and baby feeding rooms is relatively lower when compared to that of dress changing rooms in clothing outlets. Hence, it is necessary to analyze the reliability of clothing color in different scenarios. The height of the subjects is most view-invariant as reported in [6] and capturing of the attribute during subjects' appearance in a predefined region of interest in the camera view scene (entrances of private the areas in Entry-Exit surveillance scenario), makes it reliable. Similarly, build of the subjects can be captured by computing the height to width ratio provided the video footages are captured from a still camera with no variations in the view angle. Unlike the height attribute, build attribute is pose-variant. However, computing the vertical projections of the silhouettes of the segmented subjects' bounding boxes makes it discriminative. Lastly, skin complexion can be another promising attribute if illumination variations in the multiple camera views are handled.

The above discussed soft biometrics are unique in nature and are to be given equal weight for their discriminative ability under different scenarios. Also, as one

soft biometric attribute cannot single-handedly identify subjects due to intra-class variations and inter-class similarities, combinations of these attributes, to an extent, can predict the subjects by matching them with the gallery samples.

1.2 Motivations and Contributions

With the above discussion, it can be inferred that though soft biometrics are not entirely reliable to recognize subjects, they provide prominent clues and help in narrowing down the search. Also, learning of soft biometrics is computationally faster as compared to the classical visual biometrics such as face and gait. This motivates us to explore the efficiency of soft biometrics in Entry-Exit matching.

The innovations accomplished in this paper can be summarized as follows:

- First, a set of hybrid features representation that is robust to possible appearance variations is introduced.
- An ensemble-based approach for handling heterogeneous matching results from individual soft biometrics.
- Matching analysis based on single-camera as well as two-camera based Entry-Exit surveillance model.

The proposed method is evaluated using the EnEx dataset [2] that comprises of Entry-Exit surveillance data using Single-camera and EnExX dataset [13] that also comprises of Entry-Exit surveillance data but using two field-of-view-overlapping cameras.

2 Statement of the Problem

Given the input images of subjects that are segmented from the surveillance video frames, the aim is to represent each subject with a set of highly discriminative soft biometric features such as clothing color, height, body-build, and complexion. Features from the subjects that are classified to have entered private areas are extracted and are saved in the gallery with labels. For every subject who exits from a private area, features are extracted and are matched with samples in the gallery, and the matching score is computed with each labeled subject. Certainly, 100% classification accuracy cannot be expected from rank-1 but from rank-k predictions, and the goal is to find the value of k with which the search can be narrowed down from n to k where $k<<n$.

3 Proposed Model

The proposed subject recognition model is outlined in two modules, as described below.

- Feature Representation: A novel feature representation that comprises a set of features that include clothing color as well as subjects height, body-build, and complexion is introduced, thus making the model robust to clothing changes. Detailed discussion on each individual feature types can be found in the next subsection.

– Learning and Recognition: Each feature type is analyzed for its discriminative and correlating abilities on inter-class and intra-class samples, respectively, and the transformation function that maximizes the inter-class separability and intra-class associativity is computed. Subject recognition is performed based on the collective confidence of the visual soft biometrics ensembles.

3.1 Feature Representation

A set of heterogeneous features represents the individual subject, as discussed below.

Clothing Color. The given input image of the subject is decomposed into the head, torso, and leg regions based on [7]. Torso and leg regions are individually analyzed. Low-level features such as color and texture used in conventional person re-identification methods exhibit high discriminative ability on different camera views. The computational complexity of color features is faster as compared to texture features [8], and hence the proposed model is confined to color features.

RGB color space is robust to variations in translation, rotation, and scaling factors but suffers from temporal light changes outdoor throughout the day as well as illumination variations due to different camera views. HSV color space is interpreted as color space commonly perceived by humans and is considered efficient enough for color-based analysis, the RGB to HSV color transformation is time-consuming and slows down learning. On the other hand, the YCbCr color space clearly differentiates the luminance and color components. The Y component corresponds to luminance, and Cb and Cr components correspond to chrominance where Cb is blue-difference chrominance, and Cr is red-difference chrominance, and the computational complexity of RGB to YCbCr transformation is of the order $h\ X\ w$ where h and w are height and width of the input subject image. Histograms of Cb and Cr components of the torso and leg regions are computed for each subject and are concatenated to a single feature vector.

Height. Height is one of the prominent features that helps in narrowing down the search by clustering subjects of similar heights. The height of the subject is captured with reference to the entrance of private areas. As the features of the subjects who enter private areas are saved in the gallery only on confirmation of his/her entry, learning is done in time. Also, during exit, the subjects have to cross the entrance first, and hence, height is captured with other sets of features. As the height feature captured is the relative height and not the actual height, it is normalized to the values between 0 and 1.

Body-Build. Extracting height features near to the entrances of private areas also provide scope for extracting build of the subject by computing the maximum height to width ratio of the bounding boxes of the subjects when they cross the

entrance. Variations in the pose of the subject are the major challenge to be addressed. Vertical projection profile is computed for the bounding boxes, and threshold t is determined with experiments to eliminate hand and leg swings, thus segmenting the subject image based on torso distribution.

Skin Complexion. Skin complexion is another essential visual attribute that helps in grouping subjects with the same complexion. The segmented head region is analyzed for skin components using the YCbCr color model with the threshold for skin detection, as reported in [9]. The skin region is segmented from the front, lateral and oblique views except for the back view.

3.2 Matching

For each feature type, Linear Discriminant Analysis (LDA) is applied for data disassociation, by projecting the feature matrix onto a subspace that maximizes the ratio of inter-class to intra-class distribution, using the Fisher's criterion.

For every class C_i (individual), the separability d_i between samples $s \in C_i$ is computed using

$$d_i = \sum (s - \bar{s_i})(s - \bar{s_i})' \tag{1}$$

where $\bar{s_i}$ is the mean of the class C_i
The intra-class separability matrix d for n classes is given by

$$d = \sum_{i=1}^{n} d_i \tag{2}$$

and the intra-class compactness is given by

$$Q = 1/d \tag{3}$$

The disassociativity matrix D between classes is computed using

$$D = \sum_{i=1}^{n} m_i(\bar{s_i} - \bar{s})(\bar{s_i} - \bar{s})' \tag{4}$$

where m_i is the number of training samples in class C_i and $\bar{s}$ is the overall mean.

The transition matrix T that maximizes the spread between the classes and compactness within the classes is given by

$$T = |D||Q| \tag{5}$$

So, given a probe subject p, the features are extracted and every soft biometric is operated with transition function in Eq. 5 and the classification is performed for every soft biometric by computing the euclidean distance of the probe with every gallery sample and thus all the gallery classes are ranked for every soft biometric. Subject recognition is done based on the collective voting decision among the different soft biometrics ensemble thus giving equal prominence for

each soft biometric. Confidence Cf is computed for every gallery class C_i for each soft biometric f using

$$Cf(C_i, f) = \frac{n - rank(C_i, f) + 1}{n} \tag{6}$$

Hence, the Collective confidence CF of the model for each gallery class C_i is

$$CF(C_i) = \sum_{j=1}^{f} \frac{Cf(C_i, j)}{f} \tag{7}$$

Rank for every gallery class is assigned based on the collective confidence calculated using Eq. 7.

4 Experiments and Discussions

The proposed method is tested using the EnEx dataset as well as EnExX dataset. EnEx dataset provides EES data using a single camera, and EnExX dataset provides EES data using two cameras. However, EnExX dataset can be used for single camera-based analysis as well. Firstly, with a single camera view where the camera is placed so as to view the entrance of private areas without intruding the privacy of the public. Here, the camera captures flipped views of subjects during entry and exit where generally back view of the subject is visible during entry and front view during exit, or if the camera is placed so as to have a lateral view of the subject, then entry-exit shall have left-right flipped view variations. Next, with two view-overlapping cameras so as to have 360^o view of the subject where one camera compliment the other by having a flipped view of the subject simultaneously.

4.1 Experimental Setup

The input to the system is the images of the subjects extracted from video frames of the datasets using [10] deeply learned people detection method. The frames are background subtracted before applying the people detectors. Two sets of images - gallery and probes. Gallery set contains images captured while subjects entered, whereas probe set contained images captured while subjects exited. Figure 1 shows sample images of subjects in the gallery and probe using a single camera, and Fig. 2 shows sample images of subjects in the gallery and probe using two cameras. The image pair in the first row shows the entry of the subject captured in two different cameras and the second row shows subject exiting the private area

The height and body-build attributes are captured when the subject crossed the entrances of private areas, and other attributes such as clothing color and skin complexion were extracted after normalizing the input images to the standard size of 128×64. Then the input images are decomposed into the head, torso, and leg regions, and these are converted to the YCbCr color model, and histograms

Fig. 1. Sample Entry-Exit image pair with variation in appearance from EnExX dataset [13] using Single Camera

Fig. 2. Sample Entry-Exit image pair with variation in appearance from EnExX dataset captured using two cameras

of Cb and Cr components are extracted with 24 bins per channel for each region of a subject. Skin regions are extracted from each input image, and the mean of Cb and Cr components are computed.

Initially, simulations were carried out on the EnEx dataset, here the size of the private area, which is the gallery, was assumed to be 10, 25, and 50, respectively. The gallery contained 30 training images for each subject, among which five images provided height and body-build metrics. Hence, the height and body-build attributes are trained separately. For every probe subject, a single image that contained all the attributes was selected to match with all the samples in the gallery. Later, the simulations were extended to EnExX dataset, where pair of images for clothing and skin complexion attributes are provided

for training and testing, and the features are concatenated to a single vector for representation.

The notion of Entry-Exit surveillance is novel and the proposed model is first of its kind that solved Entry-Exit matching which involved appearance invariant person re identification. However, comparative analysis is provided with [11] and [12] methods that includes soft biometrics by evaluating them on the EnEx and EnExX dataset. Tables 1, 2, 3 and 4 provide matching scores of the proposed method in comparison with [11] and [12].

Table 1. Matching rates on EnEx dataset, gallery size = 25

Gallery size = 25			
Rank	1	5	10
RS-KISS	0.079	0.316	0.486
Ensemble Learning	0.094	0.367	0.491
Proposed model	0.231	0.489	0.867

Table 2. Matching rates on EnEx dataset, gallery size = 50

Gallery size = 50			
Rank	1	5	10
RS-KISS	0.071	0.283	0.423
Ensemble Learning	0.082	0.326	0.463
Proposed model	0.217	0.466	0.812

4.2 Discussion

With the results, it is evident that the state of the art person re-identification systems suffers due to variations in the clothing of the subjects and hence the demand for active research in appearance-invariant person re-identification. Variations in the appearances account for clothing, height due to change in foot wears and complexion due to applying cosmetics. Hence, collective confidence of the attributes are considered than weighted averaging. The proposed model, though, is robust to clothing variations, it suffers due to uniformity in height, build and complexion of subjects of the same race. Also, the extraction of skin attributes from subjects who cover the entire body with religious attires is challenging. The height attribute though, looks promising, it suffers due to variations in head accessories of the subjects. Overall, the evaluation results of the proposed model on the EES specific datasets can be used as a benchmark by the research community to compare their works on the EES matching problem using visual soft biometrics.

Table 3. Matching rates on EnExX dataset, gallery size = 25

Gallery size = 25			
Rank	1	5	10
RS-KISS	0.091	0.352	0.533
Ensemble Learning	0.162	0.431	0.616
Proposed model	0.366	0.581	0.891

Table 4. Matching rates on EnExX dataset, gallery size = 50

Gallery size = 50			
Rank	1	5	10
RS-KISS	0.088	0.263	0.411
Ensemble Learning	0.113	0.412	0.493
Proposed model	0.342	0.563	0.838

5 Conclusion

With this paper, it can be inferred that in real-time tracking, it is important to narrow down the search based on predictions using visual attributes whose learning and recognition are faster than the reliable attributes such as gait whose learning and recognition rates are higher and faster with the narrow galleries. Collective confidence based matching provides equal weightage to all the soft biometrics considered. The proposed novel Entry-Exit subject matching method shows good rank-10 accuracy on the standard datasets and thus enkindle competitive research in the Entry-Exit surveillance domain.

Acknowledgements. This work has been supported by The University Grants Commission, India.

References

1. Sulman, N., Sanocki, T., Goldgof, D., Kasturi, R.: How effective is human video surveillance performance? In: 19th International Conference on Pattern Recognition, ICPR 2008, pp. 1–3, 8–11 December 2008
2. Vinay Kumar, V., Nagabhushan, P., Roopa, S.N.: Entry–exit video surveillance: a benchmark dataset. In: Chaudhuri, B.B., Nakagawa, M., Khanna, P., Kumar, S. (eds.) Proceedings of 3rd International Conference on Computer Vision and Image Processing. AISC, vol. 1022, pp. 353–364. Springer, Singapore (2020). https://doi.org/10.1007/978-981-32-9088-4_30
3. Yu, H.X., Zheng, W.S., Wu, A., Guo, X., Gong, S., Lai, J.H.: Cross-view asymmetric metric learning for unsupervised person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (2019)

4. Karanam, S., Wu, Z., Rates-Borras, A., Camps, O. and Radke, R.J.: A systematic evaluation and benchmark for person re-identification: features, metrics, and datasets. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 3, pp. 523–536, 1 March 2019. https://doi.org/10.1109/TPAMI.2018.2807450
5. Chin-Poo, L., Chiat, W., Alan, T., Kian, L.: Review on vision-based gait recognition: representations, classification schemes, and datasets. Am. J. Appl. Sci. **14**, 252–266 (2017). https://doi.org/10.3844/ajassp.2017.252.266
6. Moctezuma, D., Conde, C., De Diego, I.M., et al.: J. Image Video Proc. **2015**, 28 (2015). https://doi.org/10.1186/s13640-015-0078-1
7. Iat-Fai, L., Jing-Jing, F., Ming, T.: Automatic body feature extraction from a marker-less scanned human body. Comput. Aided Des. **39**, 568–582 (2007). https://doi.org/10.1016/j.cad.2007.03.003
8. Shahbahrami, A., Borodin, D., Juurlink, B.: Comparison between color and texture features for image retrieval (2008)
9. Kolkur, S., Kalbande, D., Shimpi, P., Bapat, C., Jatakia, J.: Human skin detection using RGB. HSV, and YCbCr color models (2017). https://doi.org/10.2991/iccasp-16.2017.51
10. Kokul, T., Ramanan, A., Pinidiyaarachchi, U.A.J.: Online multi-person tracking-by-detection method using ACF and particle filter. In: 2015 IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS), Cairo, pp. 529–536 (2015)
11. Yang, Y., Liu, X., Ye, Q., Tao, D.: Ensemble learning-based person re-identification with multiple feature representations. Complexity **2018**, 12 (2018). Article ID 5940181. https://doi.org/10.1155/2018/5940181
12. Tao, D., Jin, L., Wang, Y., Yuan, Y., Li, X.: Person re-identification by regularized smoothing KISS metric learning. IEEE Trans. Circuits Syst. Video Technol. **23**(10), 1675–1685 (2013)
13. Vinay Kumar, V., Nagabhushan, P.: Monitoring of people entering and exiting private areas using computer vision. Int. J. Comput. Appl. **177**(15), 1–5 (2019)