Journal of Ambient Intelligence and Humanized Computing
https://doi.org/10.1007/s12652-022-03891-0

ORIGINAL RESEARCH

Person retrieval in surveillance videos using attribute recognition

Hiren Galiyawala[1] · **Mehul S. Raval**[1] · **Meet Patel**[2]

Received: 6 July 2021 / Accepted: 28 April 2022

Abstract
In person attribute recognition (PAR), an individual is described by his or her appearance. PAR-based person retrieval is a *cross-modal* problem where the input is a textual description of the person's appearance and the output is an image of the person. The paper introduces PAR model development by merging a large-scale RAP dataset with the person retrieval benchmark dataset of AVSS 2018 challenge II. It uses a single deep network to detect a person's attributes. The proposed approach uses five attributes; age, upper body (uBody) clothing color, uBody clothing type, lower body (lBody) clothing color, and lBody clothing type. Mask R-CNN is used for person detection, and the approach weighs each attribute to generate a ranking score for every detected person. Unlike the existing approaches, the proposed method uses a single deep network and fewer attributes to achieve state-of-the-art average intersection-of-union (IoU) of 66.7%, retrieval with IoU ≥ 0.4 is 85.6%, and an average true positive rate (TPR) of 85.30%. It is better by 10.80% average IoU, 5.94% IoU ≥ 0.4, and 3.85% TPR than the existing state-of-the-art person retrieval using attributes recognition.

Keywords Attribute weighting · Person attribute recognition · Person retrieval · Soft biometrics · Textual description

1 Introduction

Video surveillance applications are gaining attention to make the world a better and safer place to live. Person retrieval becomes an essential and critical task during an investigation using surveillance videos. Searching for a person in videos manually is an inefficient and time-consuming process. Hence, intelligent video surveillance is gathering interest within the research community.

Existing methods retrieve a person using an image as a query to the retrieval system. These systems use a large-scale cropped person image dataset created from non-overlapping cameras. Such image-based retrieval techniques are known as person re-identification (Re-ID) (Chen et al. 2018). Re-ID requires at least one query image for re-identification of the person from other camera images. Re-ID techniques fail when a query image is not available, and only textual description from the eye-witness is available. Computer vision and natural language processing research have discovered enormous opportunities for person retrieval using textual description, e.g., *a short man with black jeans and a red-colored long sleeve shirt wearing a blue cap*. Such human description contains person attributes like height (short), gender (male), clothing color (red, black), clothing type (long sleeve), and accessories (cap). These attributes are known as *soft biometrics* (Galiyawala and Raval 2021; Galiyawala et al. 2020). Figure 1 illustrates person retrieval from a surveillance frame using the soft biometrics-based textual description.

Person attributes are now input to the person retrieval system as a textual query. The system accepts one type of data, namely, text as an input query and outputs another type of data, namely, the image of the person(s). Thus, person retrieval using a textual query is also referred to as *cross-modal* retrieval (Zhen et al. 2019). PAR (Li et al. 2015; Sudowe et al. 2015) seeks to extract attributes like clothing color, gender, age, and clothing type from the person's image. It is a challenging task due to various factors. For example, varying illumination conditions make the same

✉ Hiren Galiyawala
hirenkumar.g@ahduni.edu.in

Mehul S. Raval
mehul.raval@ahduni.edu.in

Meet Patel
patel.meet.637@ldce.ac.in

[1] School of Engineering and Applied Science, Ahmedabad University, Ahmedabad, Gujarat 380009, India

[2] L. D. College of Engineering, Ahmedabad, Gujarat 380015, India

Published online: 20 May 2022

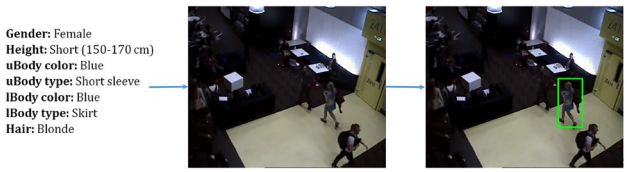


Fig. 1 Person retrieval using textual description (Halstead et al. 2018)

clothing color appear differently. Occlusion affects attributes like lBody clothing color and type. Attributes like a scarf, a backpack may be visible in one camera view and may not be visible for another camera view. Such issues in PAR make person retrieval a challenging problem.

1.1 Related work

Early research work (Jain et al. 2004) suggested the applicability of soft biometric attributes to improve the performance of primary biometric systems. Research work in Denman et al. (2009), Halstead et al. (2014), Denman et al. (2015) and Shah et al. (2017) demonstrated person retrieval with hand-crafted features before the era of deep learning. A color histogram is used in Denman et al. (2009) for clothing color feature extraction. Halstead et al. (2014) created a soft biometric attribute-based avatar, and Denman et al. (2015) extended the avatar into channel representation with histograms of oriented gradients (HoG) feature representation. Shah et al. (2017) used the ISCC-NBS color model with CIEDE2000 distance metrics for clothing color classification.

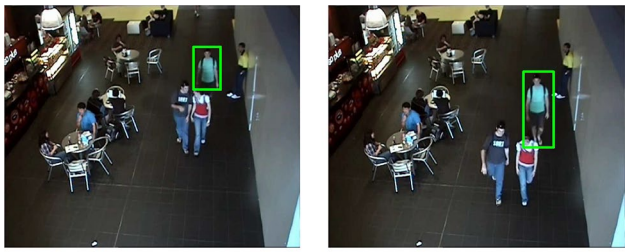
Deep convolutional neural network (DCNN) based approaches for person retrieval have been gaining prominence. Approaches in Yaguchi and Nixon (2018), Schumann et al. (2018), Galiyawala et al. (2018), Galiyawala et al. (2019), Shah et al. (2021) and Galiyawala et al. (2021) are analyzed on AVSS 2018 Challenge II dataset (Halstead et al. 2018). The challenge aims to retrieve a person using a textual query based on soft biometric attributes. Hand-crafted feature-based implementation in Denman et al. (2015) is considered as the baseline method of AVSS 2018 Challenge II (Halstead et al. 2018), and methodologies in Yaguchi and Nixon (2018), Schumann et al. (2018), Galiyawala et al. (2018), Galiyawala et al. (2019), Shah et al. (2021) and Galiyawala et al. (2021) represent deep learning-based implementations. Yaguchi and Nixon (2018) propose a transfer learning-based method for person retrieval using 9 attributes. Person detection is done using Mask R-CNN (He et al. 2017) and DenseNet-161 (Huang et al. 2017) used for attribute classification. All the attributes are predicted and the matching score is then calculated using Hamming loss. The person achieving the minimum loss is the target for a particular frame. The single shot multibox detector (SSD) is used for person detection in Schumann et al. (2018), and background modeling helps to remove false positives in person detection. It may fail to locate a non-moving person.

The methods in Galiyawala et al. (2018, 2019) and Shah et al. (2021) adopted a cascade filtering-based approach that filters out the detected person using stage-by-stage attribute filters. For example, height is the first filter, and height query is short (150–170 cm). The detected person is only available for further attribute filtering if the estimated height matches the queried height. A person is retrieved using height, clothing color, and gender in Galiyawala et al. (2018). Mask R-CNN is used for person detection. Height is estimated using the Tsai camera calibration approach (Tsai 1987) and AlexNet (Krizhevsky et al. 2012) for color and gender classification. The torso patch is extracted from the fixed torso region (20–50%), generating a noisy patch. The noisy patch leads to wrong clothing color classification.

The adaptive torso patch extraction and bounding box regression in Galiyawala et al. (2019) further improve the linear filter approach of Galiyawala et al. (2018). Galiyawala et al. (2019) proposed adaptive torso patch extraction by selecting the torso region according to the torso type attribute given in the query. Hence, color classification accuracy is improved by the removal of noisy pixels from the torso patch. Shah et al. (2021) use height, torso clothing color, torso type, torso pattern, leg clothing color, leg clothing type, leg pattern, and gender. Gender, color, and pattern classification models are developed using DenseNet-161. The ranking based approach is proposed in Galiyawala et al. (2021) and achieves state-of-the-art performance on AVSS 2018 challenge II dataset (Halstead et al. 2018) in terms of average intersection-over-union (IoU) (i.e., 0.602) and IoU ≥ 0.4 (i.e., 0.808). The approaches in Specker and Beyerer (2021); Zhao et al. (2021) investigate person retrieval using PAR based techniques on popular datasets like RAP (Li et al. 2018), but they showcase the retrieval from gallery of cropped images. Thus, challenges prevailing in full surveillance frame like detection, occlusion and varying illumination are not handled. Some recent approaches (Sakib et al. 2022; Zhao et al. 2022) of PAR showcases the state-of-the-art performance (i.e., 93.41 mA (Sakib et al. 2022)), but they do not consider clothing color attributes in recognition. However, clothing color is one of the most discriminative attributes (Galiyawala and Raval 2021) for person retrieval.

The major limitations of the approaches in Galiyawala et al. (2018, 2019, 2021) and Shah et al. (2021) are as follows:

1. Error in initial stage filters will propagate to other filters for approaches in Galiyawala et al. (2018), Galiyawala et al. (2019) and Shah et al. (2021). Figure 2a illustrates a person with partial occlusion, and the person detector prepares a smaller box in comparison with the actual



(a) (b)

Fig. 2 **a** Person with partial occlusion and **b** person without occlusion in surveillance frame (Halstead et al. 2018)

box (Fig. 2b). The smaller box leads to wrong height estimation, and if height is the first stage filter, it may remove the target person in the initial filter stage. Hence, the target person will not be available for the further filtering stages.

2. The cascade filtering approaches (Galiyawala et al. 2018, 2019; Shah et al. 2021) removes non-matching person(s) at every stage and does not consider the contribution of each queried attribute for retrieval. Hence, such an approach does not allow for weighting each soft biometric attribute for softer decision making.
3. Each soft biometric attribute requires a separate model for its recognition. Soft biometric attributes are not limited, and hence it becomes a time-consuming process for developing a new model while adding a new attribute to the in-person retrieval algorithm.
4. It becomes challenging to prepare an annotated dataset for each attribute model creation, e.g., torso pattern, shoe type.

This paper proposes PAR-based person retrieval using age, uBody clothing color, uBody clothing type, lBody clothing color, and lBody clothing type. Person detection and semantic segmentation are done using Mask R-CNN (He et al. 2017). Point-to-point multiplication is then applied between semantic segmentation (i.e., binary mask) of each detected person and the respective person image to remove the cluttered background. The model predicts the attributes with their probability score. It is then fed to the attributes score weighting model to rank the detected persons. The person with the highest score is chosen as the target. The algorithm is tested on the AVSS 2018 Challenge II dataset[1] (Halstead et al. 2018), and results are compared with the current state-of-the-art approaches.

The proposed approach overcomes above mentioned limitations of Galiyawala et al. (2018, 2019, 2021) and Shah et al. (2021) by recognizing all attributes with a *single model* and weighing them in the retrieval process. Thus, PAR model requires lesser parameters to learn compared to multiple models used in previous approaches. The state-of-the-art performance is achieved with fewer attributes. The contributions of the paper are summarized as follows:

1. A multi-attribute learning-based single model for person attribute recognition is developed. It avoids the preparation of a separate dataset and recognition model for each attribute.
2. Richly Annotated Pedestrian (RAP) (Li et al. 2018) and AVSS dataset samples are merged to cover more diversity and develop a better model. AVSS 2018 challenge II dataset (Halstead et al. 2018) does not cover detailed annotations, e.g., torso clothing type is annotated as {no sleeve, short sleeve, and long sleeve}. The RAP dataset provides finer annotations {ub-Shirt, ub-Sweater, ub-Vest, ub-TShirt, ub-Cotton, ub-Jacket, ub-SuitUp, ub-Tight, ub-ShortSleeve, ub-Others}. Finer annotations for AVSS 2018 challenge II datasets are proposed for the PAR model development.
3. The attributes score weighting model is developed to consider the contribution of each attribute during person retrieval.
4. State-of-the-art performance is achieved with fewer attributes.
5. Cropped person image gallery-based retrieval approaches do not consider challenges like pose, occlusion, and illumination in person detection from the full surveillance frame. This paper proposes PAR-based end-to-end person retrieval in surveillance videos.

Further, the paper is organized as follows. Section 2 covers the person retrieval approach, PAR model development, attribute weighting module, and person ranking strategy. Dataset preparation, implementation details for PAR model, and attribute weighting model are elaborated in Sect. 3. Experimentation results, performance analysis, comparison with current state-of-the-art approaches is discussed in Sect. 4. Section 5 concludes the paper.

2 Person retrieval approach

2.1 Overall strategy

The flow diagram of the PAR based person retrieval approach is depicted in Fig. 3. The approach uses age, uBody clothing color, uBody clothing type, lBody clothing color, and lBody clothing type. Person detection, semantic segmentation, and instance segmentation in each surveillance frame are done using Mask R-CNN (He et al.

[1] https://github.com/simondenman/SemanticSearchChallengeAVSS18.

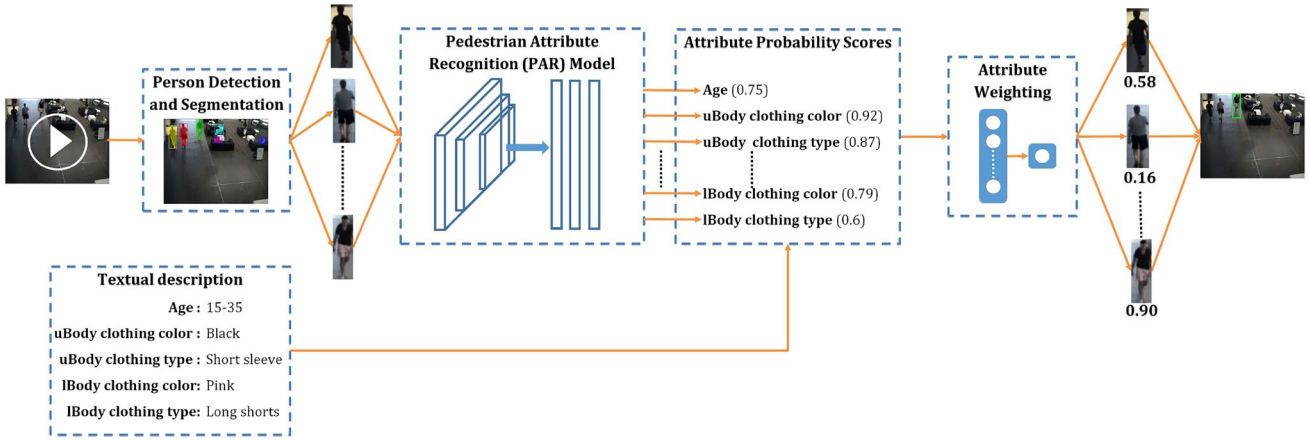


Fig. 3 PAR-based person retrieval approach. The proposed approach predicts each attribute using PAR model for all detected person in the frame. The attribute probability score vector for each person is derived based on the query attributes. These probability score vectors are then fed to attribute weighting model to generate the ranking score for each person. The person with highest ranking score is considered as the retrieved person for the given attribute query

2017). The detected person image and its corresponding segmentation mask are used in the PAR model as shown in Fig. 4, which predicts the soft biometric attribute for each detected person. It provides probability scores for each attribute. The probability scores of the query attributes are considered for each detected person. For the above example, if the age query attribute is ageLess30 and the PAR model generates a probability score of 0.75, then the age probability score is considered to be 0.75.

Similarly, the probability scores of the query attributes are derived, and score vectors are prepared for each person (Fig. 3). These probability score vectors are then fed to the attributes score weighting model to generate each person's ranking score. The person with the highest-ranking score is considered as the retrieved person for the given attribute query. However, soft biometric attributes are not unique to an individual (Galiyawala and Raval 2021). Hence, in some cases, multiple people may match the query attributes in a single frame with a similar score. In such cases, the top-2

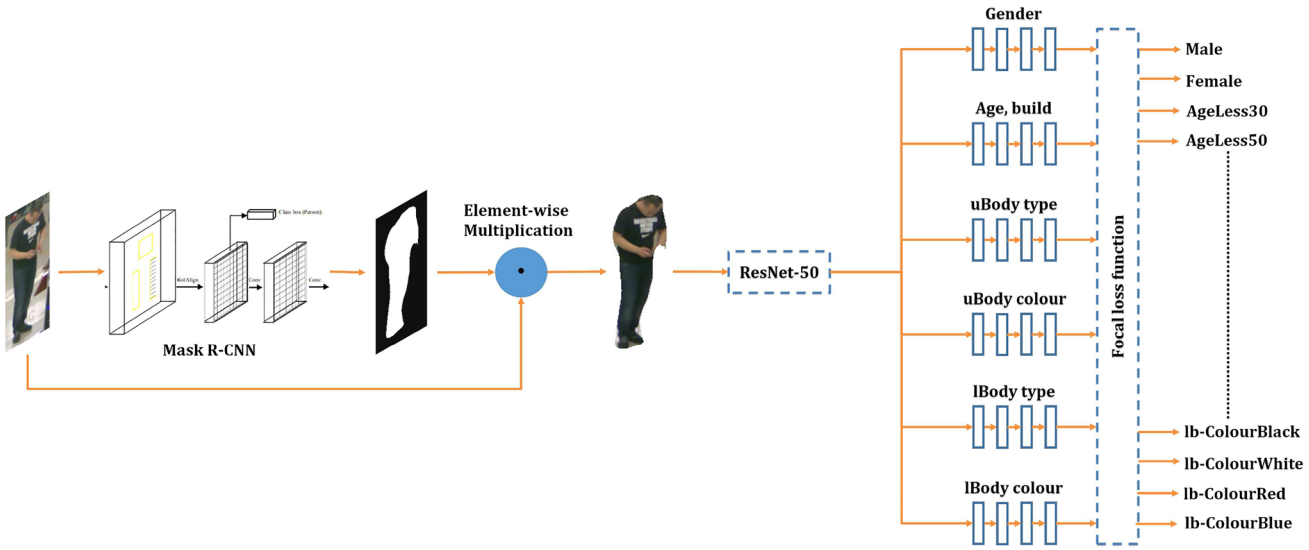


Fig. 4 Person attribute recognition model. Mask R-CNN generates the binary mask of the person. The cluttered background is removed by element-wise multiplication of person mask and original image. This enable network to focus only on person relevant information and avoids the contribution of cluttered background. ResNet-50 provides the person feature vector for multi-attribute learning where attributes are learned in separate channel of fully connected layers. Attributes are retrieved with its probability scores. Abbreviations: uBody = upper body, lBody = lower body

Fig. 5 Visualization of heat maps resulting of ResNet-50 convolutional layers using clutter free person images

persons with highest scores are considered and the person with highest IoU is declared as the target person.

Following aspects are yet unexplored in cascade filter-based approaches:

1. Cascade filter-based approaches (Galiyawala et al. 2018, 2019; Shah et al. 2021) follow the stage-wise attribute filtering. However, neither of them explored the importance of ordering the filters nor giving insight into their arrangement.
2. The number of permutations increases with a rise in attributes.
3. Finding of the optimal filter arrangement becomes computationally expensive.

PAR-based person retrieval removes such limitations by recognizing all attributes with a *single model* and weighing them in the retrieval process.

2.2 Person attribute recognition model

Figure 4 provides an overview of the PAR model architecture and the model infers the set of a person's attributes at once. It avoids the requirement of an attribute-wise separate model used in cascaded filter-based approaches (Galiyawala et al. 2018, 2019, 2021; Shah et al. 2021). The PAR model consists of four tasks: (1) semantic segmentation to generate a person mask; (2) element-wise multiplication of a person image and mask; (3) feature extraction; and (4) attribute-oriented channels for attribute learning.

The person image is first given as input to Mask R-CNN (He et al. 2017). It generates the person's binary mask. Element-wise, multiplication is done between the person image and the binary mask. It helps to remove the cluttered background from the person's image. Thus, background noise does not contribute to further feature extraction processes. Feature extraction is done using ResNet-50 (He et al. 2016). The element-wise multiplication block provides hard attention and enables the network to focus on person-relevant foreground features. The heat maps resulting from ResNet-50 convolutional layers are shown in Fig. 5. They indicate that attention is generated on the person's foreground rather than the background information.

The final task is to learn attributes in a separate channel, as shown in Fig. 4. Soft biometric attributes are labeled with multiple classes, e.g., uBody clothing type having classes like {ub-Shirt, ub-Sweater, ub-Vest, ub-TShirt, ub-Cotton, ub-Jacket, ub-SuitUp, ub-Tight, ub-ShortSleeve, ub-Others} in the RAP dataset (Li et al. 2018). The short-sleeve shirt carries two classes in uBody clothing type, i.e., ub-Shirt and ub-ShortSleeve. Learning with attribute-oriented separate channels considers the correlation among such attributes. Six attribute channels (gender, age and build, uBody type, uBody color, lBody type, and lBody color) are considered for attribute learning. It should be noted that a person's build is highly correlated with age, and hence, learning for both these attributes is done in the same channel.

Each attribute-wise channel consists of 4 fully connected layers, and implementation details are further discussed in Sect. 3.2. Each class in the attribute is assigned the weight during learning to handle the class imbalance issue. The class weight formula is as follows:

$$w_i = \frac{n_{samples}}{n_{classes} \times n_{samples_i}} \tag{1}$$

where, w_i = weight for class-i, $n_{samples}$ = total samples for each class in attribute , $n_{classes}$ = total classes in attribute, $n_{samples_i}$ = total samples in ith class of the attribute.

This strategy assigns a higher weight to the class with fewer samples and lower weight to the class with more samples in an attribute.

2.3 Attribute weighting module and person ranking strategy

The PAR model provides the person attributes with its probability scores. Let the set of attribute recognition for person image, I, be $\{a_1, a_2, \cdots, a_n\}$, where $a_i \in [0, 1]$ is the ith attribute recognition score from the PAR model. The probability scores of the query attributes are concatenated as a vector $a \in R^{1\times n}$:

$$\begin{aligned} a = \{&score_{age}, score_{uBodytype}, score_{uBodycolor}, \\ &score_{lBodytype}, score_{lBodycolor}\} \end{aligned} \tag{2}$$

A shallow neural network is trained using such vectors, and an attribute weighting model is prepared. Each frame in AVSS 2018 challenge II dataset consists of one person with ground truth data along with other persons. Now considering binary classification, the score vector for a ground

truth is labeled as '1'and score vectors for other persons are labeled as '0'. Such score vectors trains the attribute weighting model as discussed in Sect. 3.3. Hence, weights to each attribute are learned by attribute weighting model and considers the contribution of each attribute in the retrieval process. The ranking score is then learned as,

$$r = sigmoid(wa^T + b) \tag{3}$$

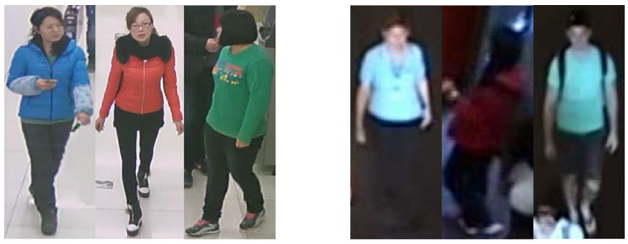
where, $w \in R^{n \times n}$ and $b \in R^{n \times 1}$ are trainable parameters. During the testing phase, score vector of each detected person is applied to the attribute weighting model which generates their *ranking* score. For example, 3 persons are detected in the video frame. Let, a_1, a_2, and a_3 are the attribute score vectors generated by PAR model. The attribute weighting model generates score r_1, r_2, and r_3 for a_1, a_2, and a_3 respectively. The person with highest score is considered as target. Since the soft biometric attributes are not unique to an individual (Galiyawala and Raval 2021), multiple people may match the query attributes in a single frame with a similar score. Hence, the proposed approach considers a rank-2 search to declare the target person. IoU is calculated in top-2 matches from a ranking score, and the person with the highest IoU score is declared as the target person.

3 Dataset and implementation details

This section covers details about the dataset, PAR model training, and attribute weighting model training. Mask R-CNN (He et al. 2017) is trained on the MS COCO dataset (Galiyawala et al. 2021) for only the person class. Weights of ResNet-50 trained on the ImageNet (Galiyawala et al. 2021) dataset are used for feature extraction in the PAR model.

3.1 Dataset and annotations

The proposed approach uses the benchmark AVSS 2018 challenge II task-2 dataset (Halstead et al. 2018) for PAR-based person retrieval. Video sequences are captured using six stationary calibrated cameras with a resolution of 704×576. The dataset consists of unconstrained video sequences of 110 persons for training and 41 persons for testing. Each training sequence is annotated with nine body markers and 16 soft biometric attributes, while the testing sequences are provided with only soft biometric attributes. Since the PAR model requires cropped person images (Fig. 4) for attribute recognition, the AVSS 2018 challenge II task-2 dataset's full surveillance frames are not directly suitable for training the PAR model. This paper uses the most popular and by far the largest RAP dataset (Li et al. 2018) and the AVSS dataset for PAR model development. The RAP dataset contains 84,928 person images collected from 25 HD (1280×720) cameras



(a) RAP dataset images (b) AVSS 2018 challenge II dataset

Fig. 6 Sample images from RAP (Li et al. 2018) and AVSS 2018 challenge II (Halstead et al. 2018) dataset

at an indoor shopping mall. The resulting images vary in size from 33×81 to 415×583.

The PAR model developed only from the RAP dataset may not achieve good performance on AVSS 2018 challenge II datasets for person retrieval from a full surveillance frame. Figure 6 shows sample images from the RAP and AVSS 2018 challenge II datasets. It should be noted that the images in Fig. 6b are cropped from full surveillance frames of the AVSS 2018 challenge II datasets. The differences between the two datasets are illumination conditions, pose, color, and view. For example, the uBody colors from left to right are blue, red, and green in both the RAP and AVSS images. However, colors in the images are perceived as entirely different due to different cameras, illumination conditions, and environments. Also, the blue and green colors appear almost similar in AVSS sample images. Moreover, the RAP dataset images have appropriate resolution and illumination conditions compared to AVSS 2018 challenge II dataset.

Thus, a single dataset may not cover such diversity and different challenges. Hence, PAR model development has to be done by merging the RAP and AVSS datasets. Nevertheless, attribute annotations are different in both datasets. For example, uBody clothing type in RAP images is annotated with ten classes, {ub-Shirt, ub-Sweater, ub-Vest, ub-TShirt, ub-Cotton, ub-Jacket, ub-SuitUp, ub-Tight, ub-ShortSleeve, ub-Others}. In comparison, AVSS 2018 challenge II images are annotated with only three classes, {long sleeve, short sleeve, no sleeve}. These annotations need to be mapped to merge the datasets.

This paper creates a dataset to develop a robust PAR model by merging RAP and AVSS images with suitable annotation mapping. The RAP dataset consists of 2589 different persons, and AVSS dataset consists of 151 different persons, including training and testing sets. AVSS 2018 challenge II dataset contains fewer classes for clothing type compared to the RAP dataset. Hence, the AVSS images are annotated as per the RAP dataset annotations of attributes and their respective classes. Both datasets have the same annotations for gender, {male, female}. Body build annotation of AVSS dataset {very large, large, average, slim, very

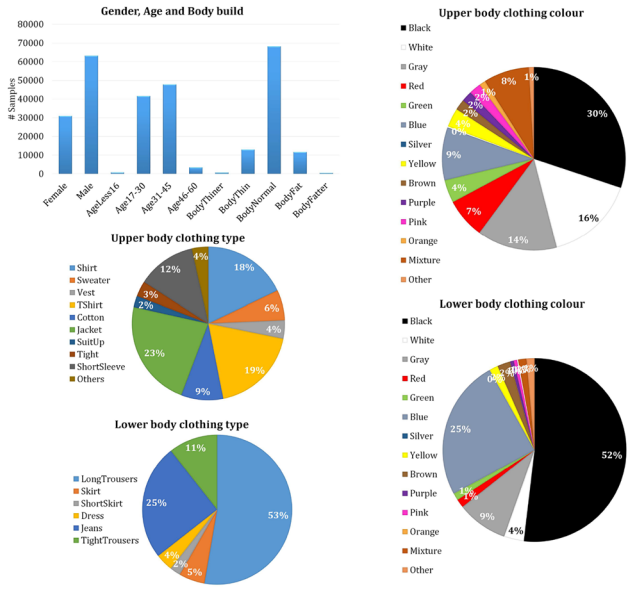


Fig. 7 The distribution of annotation on RAP (Li et al. 2018) and AVSS 2018 challenge II (Halstead et al. 2018) dataset for PAR model

slim} is mapped to {BodyFatter, BodyFat, BodyNormal, BodyThin, BodyThinner} respectively. Age annotations of AVSS dataset {< 20, 15–35, 25–45, 35–55} are mapped to {AgeLess16, Age17–30, Age31–45, Age46–60}, respectively. 'Age>50'class of AVSS dataset is not considered in the mapping due to very few samples to avoid an extreme class imbalance issue. Both datasets have the same uBody and lBody clothing color, i.e. {black, white, grey, red, green, blue, yellow, brown, purple, pink, orange}. 'Silver'color is not available in the AVSS dataset, and hence it is not considered for mapping. 'Skin'color is not available in the RAP dataset, and hence, it is mapped to color class 'other'of the RAP dataset. The uBody clothing type contains ten classes in RAP while only three classes in the AVSS dataset. Similarly, lBody clothing type contains six classes {lb-LongTrousers, lb-Skirt, lb-ShortSkirt, lb-Dress, lb-Jeans, lb-TightTrousers} in RAP while five classes Long Pants, Dress, Skirt, Long Shorts, Short Shorts in the AVSS dataset. Clothing type is not easily mapped like other attributes. Hence, they are manually annotated by observing AVSS person images.

The distribution of annotations on the combined (RAP + AVSS) dataset is shown in Fig. 7. Almost 67% of samples are male for the gender attribute. Age17–30 and Age31–45 classes cover almost 95% of samples for the age attribute, and BodyNormal covers 72% of samples in the build attribute. 60% of samples are covered by {ub-Shirt, ub-TShirt, ub-Jacket} in uBody clothing type and {lb-LongTrousers, lb-Jeans} cover 78% of samples in lBody clothing type. More than 50% of samples are of {black, white, blue} colors for uBody and lBody clothing. Such distribution shows class imbalance in the dataset, and it is overcome by assigning weights to each class during training as discussed in Sect. 2.2 (refer 1). Let us consider a gender, {male, female} attribute with $n_{samples} = 75,603$, $n_{classes} = 2$, $n_{samples_{Female}}$=24,832, and $n_{samples_{Male}}$=50,771. By considering 1, the class weights are $w_{Female} = 1.52$, and $w_{Male} = 0.74$. It indicates that *female* class is assigned a higher weight compared to *male* class to handle class imbalance issue. It is significant to note that by far, the PAR implementations (Sakib et al. 2022; Zhao et al. 2022; Li et al. 2015, 2018) do not consider clothing color attributes in recognition (although color annotations are provided). However, it is one of the most discriminative attributes (Galiyawala et al. 2018, 2019) for person retrieval. Hence, this paper, for the first time, considers the clothing color attribute in the PAR model development and shows how this attribute can be used effectively for person retrieval.

3.2 PAR model training

A ResNet-50 based feature vector from person image is now learned in 6 attribute channels. Each attribute channel consists of 4 fully connected layers with 1024, 512, 256, and 64 units, respectively, with a 'relu 'activation function. Usually, multi-attribute classification is done with a sigmoid activation function (to produce an output between 0 and 1) and binary cross-entropy (BCE) loss function. The dataset of person attribute recognition consists of a significant difference in the number of samples from different classes. In such a scenario, BCE loss function gradients are dominated by the attribute with large samples. Such an issue is handled by the weighted binary cross-entropy (WBCE) loss function proposed in Li et al. (2015). The WBCE loss considers the class imbalance problem, but it does not consider the difficulty of classifying the sample. Also, it does not classify complex samples correctly, and the network leans towards the simple samples (Lin et al. 2017). Hence, the proposed approach adapts to the sigmoid activation function with focal loss at the final classification layer. The focal loss modulates BCE and learns hard examples easily and efficiently. The output layer for each attribute consists of 1 unit with a 'sigmoid'activation function. Table 1 shows parameter settings for the PAR model training.

3.3 Attribute weighting model training

Figure 8 shows the shallow neural network, which predicts the ranking score of each detected person. It takes a probability score vector $a \in R^{1\times n}$ (output of PAR model) as input . The network consists of two hidden layers (with 128 and 64 units respectively and a 'relu'activation function) and an output layer (with 1 unit and a 'sigmoid'activation function). The network is trained for 10 epochs and an Adam optimizer is used with a learning_rate = 0.001, epsilon = 1e-07, beta_1

Table 1 Parameter setting for PAR model training

Parameter	Value
Number of images	94,184 (84,217(RAP) + 9967 (AVSS))
Training images	75,603 (67,368(RAP) + 8235 (AVSS))
Testing images	18,581 (16,849(RAP) + 1732 (AVSS))
Image input shape	224 × 224 × 3
Learning rate	0.001
Drop-out probability	0.4
Batch size	32
Number of epochs	9
Optimizer	Stochastic Gradient Descent (SGD)
Weight decay	0.0005
Momentum	0.9

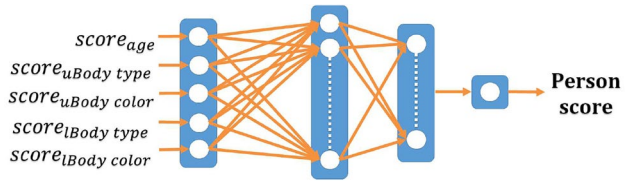


Fig. 8 Attribute weighting model and person score generation

= 0.9 and beta_2 = 0.999. The training data of the attribute score vectors are prepared from the AVSS challenge II (Halstead et al. 2018) training dataset. 12,284 attribute score vectors are prepared. The ground truth person in the frame is labeled as '1', and other persons are labeled '0'.

4 Experimentation results

The experimentation results are derived on AVSS 2018 challenge II (Halstead et al. 2018) (task-2) test dataset. The proposed approach results are compared with the baseline (Denman et al. 2015) method (avatar-based) and current state-of-the-art methods (CNN-based) of Yaguchi and Nixon (2018), Schumann et al. (2018), Galiyawala et al. (2018, 2019, 2021) and Shah et al. (2021). IoU and true positive rate (TPR) metrics are used for performance evaluation of the proposed approach. The person localization accuracy is measured by IoU and it is given by:

$$IOU = \frac{GT \cap D}{GT \cup D} \tag{4}$$

where, D = bounding box output of the algorithm and GT = ground truth bounding box. IoU is an evaluation metric used to measure accuracy of person detection in the given dataset. The proposed model generates bounding box at the output and therefore can be evaluated using IoU. It provides an idea about how accurate the algorithm is in localizing person(s) compared to the ground truth.

TPR for the person retrieval is calculated as:

$$TPR(\%) = \frac{Number\ of\ frames\ with\ correct\ retrieval}{Total\ frames} \times 100 \tag{5}$$

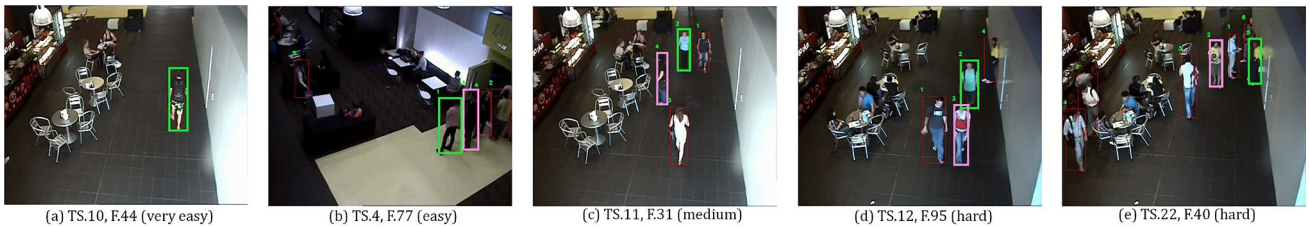
The performance comparison is made for average IoU, percentage retrieval with IoU ≥ 0.4, TPR (%), and the number of soft biometrics used for person retrieval. Table 2 shows each attribute accuracy achieved by the PAR model (Sect. 2.2) on the combined RAP and AVSS datasets. The average accuracy of the PAR model is 92.06%.

Table 2 Each soft biometric attribute accuracy achieved by PAR model

Attribute	Accuracy (%)	Attribute	Accuracy (%)
Female	90.31	ub-ColorYellow	96.40
Male	90.29	ub-ColorBrown	97.33
AgeLess16	99.28	ub-ColorPurple	97.98
Age17–30	64.29	ub-ColorPink	96.95
Age31–45	62.52	ub-ColorOrange	98.15
Age46–60	96.66	ub-ColorMixture	91.93
BodyThiner	99.16	ub-ColorOther	98.56
BodyThin	89.35	lb-LongTrousers	80.20
BodyNormal	76.18	lb-Skirt	96.20
BodyFat	87.33	lb-ShortSkirt	97.65
BodyFatter	99.47	lb-Dress	96.77
ub-Shirt	87.19	lb-Jeans	86.87
ub-Sweater	92.38	lb-TightTrousers	93.16
ub-Vest	97.02	lb-ColorBlack	82.66
ub-TShirt	79.18	lb-ColorWhite	97.16
ub-Cotton	90.96	lb-ColorGray	90.34
ub-Jacket	78.37	lb-ColorRed	98.26
ub-SuitUp	97.62	lb-ColorGreen	98.57
ub-Tight	95.9	lb-ColorBlue	88.36
ub-ShortSleeve	92.17	lb-ColorSilver	99.97
ub-Others	96.96	lb-ColorYellow	98.06
ub-ColorBlack	78.83	lb-ColorBrown	97.95
ub-ColorWhite	82.37	lb-ColorPurple	99.70
ub-ColorGray	83.52	lb-ColorPink	99.20
up-ColorRed	92.33	lb-ColorOrange	99.79
ub-ColorGreen	95.14	lb-ColorMixture	98.37
ub-ColorBlue	93.15	lb-ColorOther	99.02
ub-ColorSilver	99.72	**Average**	**92.06**

4.1 Qualitative results

Figure 9 shows the sample frames where the person is retrieved correctly. The abbreviation: TS.10, F.44 (very easy) indicates Test Sequence 10 with frame number 44



(a) TS.10, F.44 (very easy) (b) TS.4, F.77 (easy) (c) TS.11, F.31 (medium) (d) TS.12, F.95 (hard) (e) TS.22, F.40 (hard)

Fig. 9 True positive cases of person retrieval using textual description. Abbreviation: TS.10, F.44 (very easy) indicates Test Sequence 10 with frame number 44 and very easy level of difficulty in the test dataset. Person with Mask R-CNN detection score < 0.35 is not considered in retrieval process. The 'green'box indicates rank-1 person and 'pink 'box indicates rank-2 person in the frame. Person(s) in retrieval process are shown with numbers for better understanding

and a very easy difficulty level in the test dataset. The person with a detection score < 0.35 is not considered in the person retrieval process. The AVSS 2018 challenge II test dataset provides uBody clothing color annotations as {torso color-1, torso color-2} and similarly for lBody clothing color {leg color-1, leg color-2}. In the case of two-color annotations, the highest score color is considered for retrieval. The 'green'box indicates a rank-1 person in each sample frame, and the 'pink 'box indicates a rank-2 person in the frame. The person(s) in the retrieval process is shown with numbers for further better discussions.

Figure 9a shows a person from TS.10, F.44 with a very easy difficulty level where the target person is visible, illumination condition is fair, and the frame has no crowd. The textual description to the system is {age: Age17-30, uBody color: ub-ColorBlack, uBody type: ub-ShortSleeve, lBody color: lb-ColorGreen, lBody type: lb-Skirt}. The person is retrieved with rank-1. The person (TS.4, F.77) with an easy difficulty level and textual description {age: Age17-30, uBody color: ub-ColorPink, uBody type: ub-ShortSleeve, lBody color: lb-ColorBlack, lBody type: lb-Jeans} is shown in Fig. 9b. The scene contains many persons, but they do not occlude the target person. The attribute weighting model generates the ranking score as {0.9362, 0.0019, 0.0030, 0.0023, 0.0024} corresponding to five detected persons in the frame. Person-1 in the frame achieves a score of 0.9362 and is retrieved as a rank-1 search.

Similarly, Fig. 9c shows the target person (TS.11, F.31) retrieval with medium difficulty with rank-1 using ranking score as {0.0010, 0.9741, 0.0013, 0.0171}. The textual description is {age: Age46-60, uBody color: ub-ColorBlue, uBody type: ub-ShortSleeve, lBody color: lb-ColorBlack, lBody type: lb-Long-Trousers}. Here, person-2 achieves the highest score. Figure 9d showcases a hard difficulty level where the person (TS.12, F.95) is partially occluded, and a medium crowd is present. The textual description is {age: Age17–30, uBody color: ub-ColorGreen, uBody type: ub-ShortSleeve, lBody color: lb-ColorGray, lBody type: lb-Skirt}. 4 persons are available in the retrieval process. The ranking scores are {0.4319, 0.9607, 0.6141, 0.0206}. Person-2 with the highest score is declared as the target person with rank-1 search. This example depicts the correct retrieval in a partial occlusion scenario. The cascade filtering approaches in Galiyawala et al. (2018, 2019) and Shah et al. (2021) may remove the target person at the initial height filter, while the proposed approach allows such a person to remain in the retrieval process.

Figure 9e shows a person from TS.22, F.40 with a hard difficulty level where the scene contains a crowd and persons with similar appearances. The textual description is {age: Age17–30, uBody color: ub-ColorYellow and ub-ColorBlack, uBody type: ub-ShortSleeve, lBody color: lb-ColorBrown, lBody type: lb-Skirt}. Seven persons are available in the retrieval process, where person-2 and 5 appear similar. Ranking scores are {0.1067, 0.7628, 0.2008, 0.2467, 0.7905, 0.6257, 0.5322}. This example also showcases rank-1 (person-5) and rank-2 (person-2) persons with less discriminative scores due to a similar appearance. In such cases, the IoU metric is used to decide the target person. Person-5 achieves IoU '0', and person-2 achieves IoU '0.917'. Thus, person-2, i.e., the rank-2 person, is declared as the target. The sample result discussed in Fig. 9 depicts the solution to the limitations of Galiyawala et al. (2018, 2019) and Shah et al. (2021) discussed in Sect. 1.1.

Figure 10 shows the failure cases of person retrieval. Figure 10a shows the TS.12, F.42 in which the target person is occluded and very far from the camera. It is the same person retrieved correctly for F.95 (Fig. 9d) where the person is in the camera near the field. Figure 10b shows the person is merging with the background information, and hence person information is not retrieved correctly. Mask R-CNN fails to create a good person mask in Fig. 10c, and hence, the PAR model fails to recognize attributes correctly. Figure 10d shows the frame where many persons are available with a similar appearance. Rank-2 search-based strategy fails in such cases.

(a) TS.12, F.42 (occlusion) (b) TS.2, F.30 (background merging) (c) TS.30, F.56 (person detector) (d) TS.36, F.75 (similar appearance)

Fig. 10 Failure in person retrieval process by the proposed method

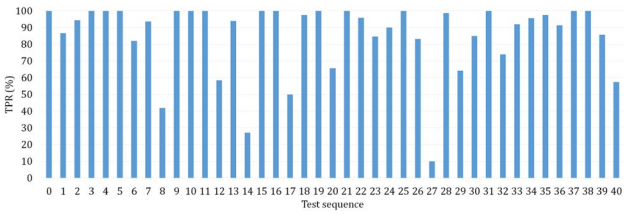


Fig. 11 TPR (%) for AVSS 2018 challenge II (task-2) test dataset

Table 3 Performance comparison with different methods on AVSS 2018 challenge II (task-2) test dataset (Halstead et al. 2018)

Methods	Average IoU	IoU ≥ 0.4	TPR (%)	#Soft biometrics
Baseline (Denman et al. 2015)	0.290	0.493	–	7
Galiyawala et al. (2018)	0.363	0.522	54.12	4
Schumann et al. (2018)	0.503	0.759	–	9
Yaguchi and Nixon (2018)	0.511	0.669	–	9
Galiyawala et al. (2019)	0.569	0.746	76.21	5
Shah et al. (2021)	0.566	0.792	–	9
Galiyawala et al. (2021)	0.602	0.808	82.14	5
Proposed (attention + focal loss)	**0.667**	**0.856**	**85.30**	**5**

4.2 Performance analysis and comparison

Figure 11 shows the TPR(%) for AVSS 2018 challenge II (task-2) test dataset (41 test sequences), and Table 3 shows the performance comparison with current state-of-the-art methods. The proposed approach achieves a state-of-the-art average TPR of 85.30%, an 3.85% improvement over 82.14% of Galiyawala et al. (2021). Among 41 persons, 32 persons are retrieved with TPR greater than 80%, 6 with TPR between 50% and 80%, 1 with TPR between 30% and 50%, and only two persons had TPR less than 30%. It indicates that more than 75% of persons achieve correct retrieval with 80% or higher TPR. It is the best performance compared to all previous approaches (Denman et al. 2015; Yaguchi and Nixon 2018; Schumann et al. 2018; Galiyawala et al. 2018, 2019, 2021; Shah et al. 2021). The proposed approach also achieves state-of-the-art performance in terms of average IoU of 0.667 and percentage of retrieval with IoU ≥ 0.4 of 0.856. It outperforms all previous approaches (Table 3) in terms of both metrics. The algorithm achieves a 10.80% improvement in average IoU and 5.94% higher IoU ≥ 0.4 than the state-of-the-art approach of Galiyawala et al. (2021). The proposed approach achieves state-of-the-art performance by using just five soft biometric attributes and a single model for attribute recognition compared to multiple models used in Galiyawala et al. (2018, 2019, 2021) and Shah et al. (2021).

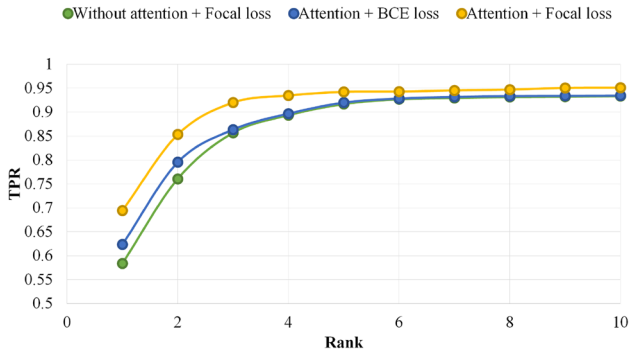


Fig. 12 Comparison of method's TPR for different ranks

Table 4 further provides the performance evaluation for attention and different loss functions. As discussed in Sect. 2.2, the PAR model is trained on images by generating hard attention with focal loss to classify complex samples correctly. The PAR model is also trained on images without attention to validate its effectiveness. As shown in Table 4, both metrics shows lower performance compared to attention based model. The experimentations are also done to analyze the performance of the focal loss over BCE loss. The results in Table 4 shows that use of focal loss with attention yields the best results. Figure 12 shows the performance of these methods in terms of TPR for rank-1 to rank-10. The method with attention and focal loss outperforms the other methods. It achieves more than 90% of TPR from rank-3 and achieves 95.09% of TPR at rank-10. The other

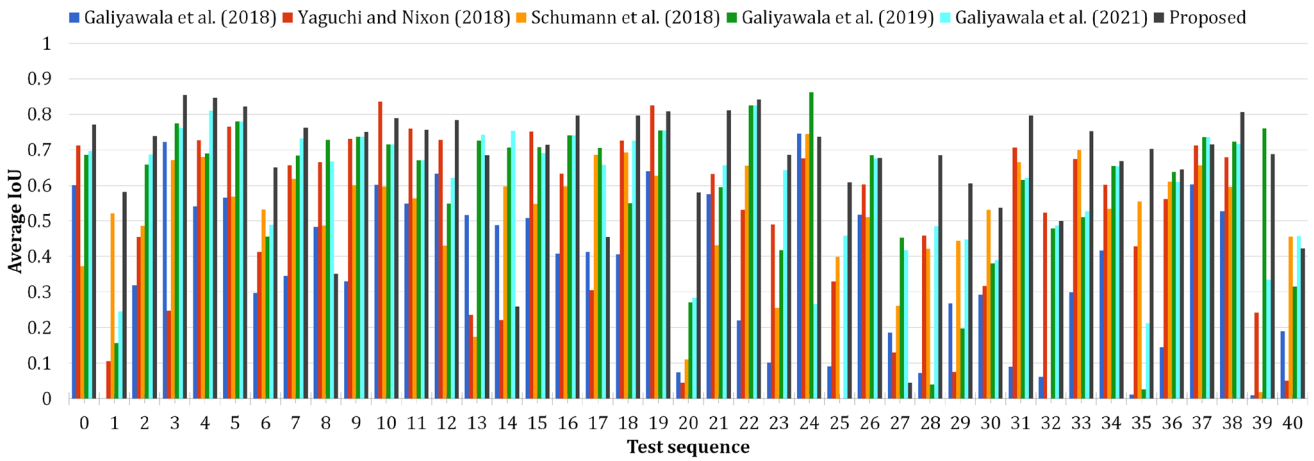


Fig. 13 Localization accuracy of each test sequence

Table 4 Performance evaluation for attention and loss functions

Approach	Average IoU	IoU ≥ 0.4
Without attention + focal loss	0.622	0.790
Attention + BCE loss	0.635	0.810
Attention + focal loss	**0.667**	**0.856**

two approaches achieve more than 90% TPR from rank-5 onwards. The diversified dataset, attention with focal loss and class weight leads to performance improvement on AVSS 2018 challenge II dataset. The comparison of person localization accuracy in terms of average IoU is shown in Fig. 13. The proposed approach achieves highest localization accuracy for 26 sequences {0, 1, 2, 3, 4, 5, 6, 7, 9, 12, 16, 18, 20, 21, 22, 23, 25, 28, 29, 30, 31, 33, 34, 35, 36, 38} out of 41.

The AVSS 2018 challenge II (Halstead et al. 2018) divides the test dataset into four difficulty levels; very easy, easy, medium, and hard. 51% of test persons represent medium and hard difficulty levels (Halstead et al. 2018; Galiyawala et al. 2019). It motivates the development of a robust algorithm. Difficulty level-wise performance comparison is shown in Fig. 14. It is evident from Fig. 14 that performance deteriorates for all approaches as the level of difficulty increases. It can be observed from that the proposed approach outperforms all previous approaches (Yaguchi and Nixon 2018; Schumann et al. 2018; Galiyawala et al. 2018, 2019, 2021) in terms of average IoU. The performance increases by 9.99%, 4.46%, 20.45%, and 15.20% for difficulty levels; very easy, easy, medium, and hard, respectively, over the state-of-the-art methods (Yaguchi and Nixon 2018; Schumann et al. 2018; Galiyawala et al. 2018, 2019, 2021) in terms of average IoU. The proposed approach performs exceptionally well for TS. 20, 21, 23, 25, 28, 29 where TS. 20, 25 represent occlusion, TS. 21 represents a change in color appearance due to illumination, TS. 23, 29 represent low illumination with crowded scenarios, and TS. 28, 29 represent the presence of persons with similar appearances. It indicates that the proposed approach performs well against such challenging scenarios.

TS. 8, 14, 27 are the sequences where the algorithm fails to achieve a good performance. In TS.8, multiple people with similar appearances are present, and the target person fails to achieve the higher ranking score. The target person in TS.14 is with two uBody clothing colors, i.e., orange and white. The scene contains more persons with white as uBody clothing color, and hence, the target person was unable to achieve the higher ranking score. The dataset distribution in Fig. 7 shows that the uBody with orange color carries only 1% of the samples. Hence, the PAR model fails to generate a higher probability score for the color. Similarly, TS.27 fails to achieve the required performance due to similar clothing types.

The proposed approach uses a single model for PAR, and hence fewer parameters are required to be learnt. The PAR model with ResNet50 (Fig. 4) requires 40.13M parameters to be learnt. While the approach in Galiyawala et al. (2018, 2019, 2021) and Shah et al. (2021) requires separate models

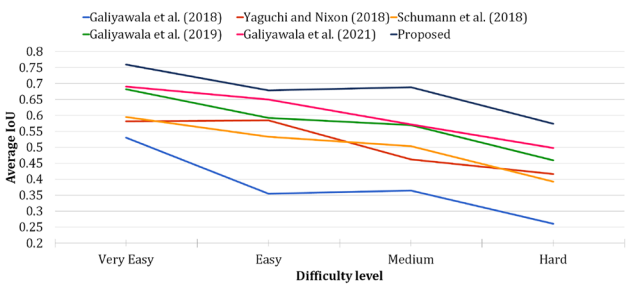


Fig. 14 Difficulty level wise performance comparison in terms of Average IoU

Table 5 Ablation experimentation with different backbone networks on AVSS 2018 challenge II (task-2) test dataset

Backbone network	Average IoU	IoU ≥ 0.4	Parameters
AlexNet	0.584	0.746	121.76M
DenseNet201	0.624	0.788	**35.85M**
Res2Net50	0.659	0.840	42.32M
ResNet50	**0.667**	**0.856**	40.13M

Bold indicates that ResNet50 backbone network performs better compared to other networks in terms of Average IOU and IoU ≥ 0.4. DenseNet201 requires fewer parameters to learn compared to other networks

for each attribute. For example, the color model (25.6M) and gender (23.6M) only together require 49.2M parameters to be learnt using ResNet50. Even more parameters will be required to be learnt if attributes are increased for the retrieval process. Thus, the proposed approach also provides cost and time-effective solutions with a perspective of real-time implementations.

4.3 Ablation study

Ablation experimentations are performed to analyze the effect of different backbone networks in the PAR model. AlexNet (Krizhevsky et al. 2012), DenseNet201 (Huang et al. 2017), Res2Net50 (Gao et al. 2019) and ResNet50 (He et al. 2016) backbone networks are studied. The attribute weighting models are also developed separately based on the respective backbone network-based PAR model's score vectors. Table 5 shows the ablation experimentations with different backbone networks. ResNet50 performs better compared to other networks, and Dense-Net201 requires fewer parameters to learn.

5 Conclusion

This paper proposes a PAR based end-to-end algorithm for person retrieval in surveillance using age, uBody clothing type, uBody clothing color, lBody clothing type, and lBody clothing color. It avoids the preparation of the dataset and the development of a separate recognition model for each attribute. The proposed approach achieves an average IoU of 0.667 and percentage of retrieval with IoU ≥ 0.4 of 0.856, surpassing the current state-of-the-art approaches by a large margin. The result also shows better performance at all difficulty levels. The proposed work does not consider gender as one of the soft biometrics attributes as the PAR model is biased more towards a male class. It indicates that the future work requires development of a better PAR model. Work can also be done on better data preparation task as the current work concentrates on annotation mappings, and not data augmentations.

Acknowledgements The authors acknowledge NVIDIA Corporation's support by way of a donation of the Quadro K5200 GPU used for this research. We would also like to thank the AVSS 2018 challenge II organizers and Mr. Dhyey Savaliya for providing inputs at various stages.

References

Chen D, Li H, Liu X, Shen Y, Shao J, Yuan Z, Wang X (2018) Improving deep visual representation for person re-identification by global and local image-language association. In: Proceedings of the European conference on computer vision (ECCV), pp 54–70,

Denman S, Fookes C, Bialkowski A, Sridharan S (2009) Soft-biometrics: unconstrained authentication in a surveillance environment. In: Proceedings 2009 digital image computing: techniques and applications (DICTA). IEEE, pp 196–203

Denman S, Halstead M, Fookes C, Sridharan S (2015) Searching for people using semantic soft biometric descriptions. Pattern Recognit Lett 68(2):306–315. https://doi.org/10.1016/j.patrec.2015.06.015

Galiyawala H, Raval MS (2021) Person retrieval in surveillance using textual query: a review. Multim Tools Appl 80(18):27343–27383. https://doi.org/10.1007/s11042-021-10983-0

Galiyawala H, Shah K, Gajjar V, Raval M S (2018) Person retrieval in surveillance video using height, color and gender. In: 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS). IEEE, pp 1–6

Galiyawala H, Raval MS, Dave S (2019) Visual appearance based person retrieval in unconstrained environment videos. Image Vis Comput. https://doi.org/10.1016/j.imavis.2019.10.002

Galiyawala H, Raval M S,, Laddha A(2020) Person retrieval in surveillance videos using deep soft biometrics. In: Richard J, Chang-Tsun L, Danny C, Weizhi M, Christophe R (eds) Deep biometrics. Springer, , pp 191–214

Galiyawala H, Raval MS, Savaliya D (2021) Dsa-pr: discrete soft biometric attribute-based person retrieval in surveillance videos. In: 2021 17th IEEE international conference on advanced video and signal based surveillance (AVSS). IEEE, pp 1–7

Gao S, Cheng M, Zhao K, Zhang X, Yang M, Torr P (2019) Res2net: a new multi-scale backbone architecture. IEEE Trans Pattern Anal Mach Intell 43(2):652–662. https://doi.org/10.1109/TPAMI.2019.2938758

Halstead M, Denman S, Sridharan S, Fookes C (2014) Locating people in video from semantic descriptions: a new database and approach. In: 2014 22nd international conference on pattern recognition (ICPR). IEEE, pp 4501–4506

Halstead M, Denman S, Fookes C, Tian Y, Nixon MS (2018) Semantic person retrieval in surveillance using soft biometrics: AVSS 2018 challenge II. In: 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS). IEEE, pp 1–6

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 770–778

He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision (ICCV), pp 2961–2969

Huang G, Liu Z, Van Der Maaten L, Weinberger K (2017) Densely connected convolutional networks. In: Proceedings of the IEEE

conference on computer vision and pattern recognition (CVPR), pp 4700–4708
Jain AK, Dass SC, Nandakumar K (2004) Can soft biometric traits assist user recognition? In: Biometric technology for human identification, vol 5404, pp 561–572
Krizhevsky A, Sutskever I, Hinton G (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25
Li D, Chen X, Huang K (2015) Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In: 2015 3rd IAPR Asian conference on pattern recognition (ACPR). IEEE, pp 111–115
Li D, Zhang Z, Chen X, Huang K (2018) A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. IEEE Trans Image Process 28(4):1575–1590. https://doi.org/10.1109/TIP.2018.2878349
Lin T, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision(ICCV). IEEE, pp 2980–2988
Sakib S, Deb K, Dhar P, Kwon O (2022) A framework for pedestrian attribute recognition using deep learning. Appl Sci 12(2):622. https://doi.org/10.3390/app12020622
Schumann A, Specker A, Beyerer J (2018) Attribute-based person retrieval and search in video sequences. In: 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS). IEEE, pp 1–6
Shah P, Raval MS, Pandya S, Chaudhary S, Laddha A, Galiyawala H (2017) Description based person identification: use of clothes color and type. In: National conference on computer vision, pattern recognition, image processing, and graphics. Springer, pp 457–469
Shah P, Garg A, Gajjar V (2021) Per-vis: Person retrieval in video surveillance using semantic description. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV), pp 41–50
Specker A, Beyerer J (2021) Improving attribute-based person retrieval by using a calibrated, weighted, and distribution-based distance metric. In: 2021 IEEE international conference on image processing (ICIP). IEEE, pp 2378–2382
Sudowe P, Spitzer H, Leibe B (2015) Person attribute recognition with a jointly-trained holistic CNN model. In: Proceedings of the IEEE international conference on computer vision workshops. IEEE, pp 87–95
Tsai R (1987) A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. IEEE J Robot Autom 3(4):323–344. https://doi.org/10.1109/JRA.1987.1087109
Yaguchi T, Nixon MS (2018) Transfer learning based approach for semantic person retrieval. In: 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS). IEEE, pp 1–6
Zhao Y, Shen C, Yu X, Chen H, Gao Y, Xiong S (2021) Learning deep part-aware embedding for person retrieval. Pattern Recognit. https://doi.org/10.1016/j.patcog.2021.107938
Zhao Y, Yam G, Lu J, Bian Z, Tian J(2022) Flsrnet: pedestrian attribute recognition using focal label smoothing regularization. Signal Image Video Process. https://doi.org/10.1007/s11760-021-02099-7
Zhen L, Hu P, Wang X, Peng D (2019) Deep supervised cross-modal retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 10394–10403