Alexandria Engineering Journal (2023) 63, 441–454

Alexandria University

Alexandria Engineering Journal

www.elsevier.com/locate/aej
www.sciencedirect.com

Attribute based spatio-temporal person retrieval in video surveillance

Rasha Shoitan [a], **Mona M. Moussa** [a,*], **Heba A. El Nemr** [b,a]

[a] *Computer and Systems Department, Electronics Research Institute, Egypt*
[b] *Computer and Software Engineering, Misr University for Science and Technology, Egypt*

Received 13 February 2022; revised 10 June 2022; accepted 26 July 2022
Available online 04 August 2022

KEYWORDS

Video surveillance;
Multi-object tracking;
Person retrieval;
Attributes description

Abstract Many venues, such as airports, railway stations, and shopping malls, have video surveillance systems for security and monitoring. However, searching for and retrieving people based on attribute descriptions in a large number of videos is difficult, particularly with weather variations and crowded places. Most of the existing attribute-based person retrieval systems consist of two main modules: object detection and person attribute recognition. The common drawbacks of object detection in the existing methods are false-positive, missing detection, and multi bounding boxes for the same object. Moreover, attribute recognition algorithms suffer from low accuracy for a single attribute classifier, while attributes error spread in the cascading multi-attribute classifier. This paper overcomes these issues by applying the ByteTrack algorithm instead of object detection to exploit the person's spatio-temporal information and generate a tube that maintains all the boxes that include the objects and associates high and low score boxes of the objects without raising false positive detection. Also, linking each person bounding boxes together results in more accurate attributes recognition than defining the attributes of each bounding box separately. Moreover, the proposed algorithm merges between selected predictions of two attribute recognition algorithms to improve the recognition performance. An extensive empirical evaluation was carried out on the SoftBioSearch database. The simulation results reveal that the proposed retrieval algorithm provides effective retrieval performance that exceeds the best conventional method by 14%.

1. Introduction

During the last decade, surveillance cameras have spread quickly; and their spreading is predicted to increase rapidly in the following years. Searching manually within these vast amounts of the created surveillance videos to retrieve a suspect in crimes takes months to complete. Thus, this raises the need for an automatic person retrieval method. Some of the existing person retrieval approaches idea to retrieve the person of interest depends on the similarity between a query person image and a list of person images extracted from the surveillance video. However, these approaches require an initial person image for the suspect to search within the video, which is

* Corresponding author.
E-mail address: mona_moussa@eri.sci.eg (M.M. Moussa).
Peer review under responsibility of Faculty of Engineering, Alexandria University.

https://doi.org/10.1016/j.aej.2022.07.053
1110-0168

unavailable in most criminalities. If the witness description of a suspect is the only available information in any crime, many other approaches propose methods to retrieve the persons depending on the semantic description. Galiyawala et al. [1] retrieve the person by height, gender, and cloth color attributes. First, the method detects and segments the persons using Mask RCNN. Then, the identified persons are processed through three cascaded filtering stages: height, gender, and cloth color to retrieve only the required person. Height is estimated by the camera calibration technique, while cloth color and gender are predicted based on a fine-tuned AlexNet. The cascading filter method does not require determining all detection person attributes. However, its main problem is that if the first filter does not predict the descriptor, the error will propagate to the successive filters, affecting retrieval accuracy.

On the other hand, Schumann et al. [2] use the Single-Shot Multibox Detector (SSD) for the detection stage and a group of classifiers for predicting the detected person attributes, which are gender, pose, torso type, torso color, torso texture, luggage, leg cloth type, color and texture. The Euclidean distance is calculated between the detected person's attributes probability and the query attributes to find the best match. However, it was found that SSD generates false positive detections, distracting the person attribute classifiers and leading to false retrieval persons. Therefore, a mixture of Gaussian distributions is used to filter the detected persons according to their motion in the detection region, which removes motionless detected persons. Yaguchi et al. [3] detect the persons using the mask R-CNN for person detection while using the DenseNet-161 for attribute classification. First, all the attributes of the detected person bounding boxes are estimated in parallel, which are gender, pose, torso type, torso color, torso texture, luggage, leg type, leg color, and leg texture, and then the hamming loss is calculated to find the matching score between the query and the detected persons. The lowest loss is the required person.

Shah et al. [4] propose cascade filters for the person attributes recognition to scale down the number of detected persons and find the suspected person. First, the persons are detected and segmented using Mask-RCNN, then the segmented persons are fed into a sequence of filters in the following order: height; torso color, torso type, torso pattern, leg color, leg pattern, and gender. The person's height is estimated by the Tsai camera calibration approach [5], while fine-tuned DenseNet-161 is used to predict the torso's color, type, and pattern of leg and gender. Galiyawala et al. [6] improve their cascading filter method in [1] by presenting an adaptive torso patch extraction and IoU-based bounding box regression to increase the retrieval performance. In addition, fine-tuned Mask R-CNN and DenseNet-169 are used for detection and attribute recognition, respectively.

As noticed earlier, the object detection algorithm is considered the backbone of the person retrieval algorithms; if the object detection fails to localize the persons; the performance of the retrieval algorithm will be affected. SSD and Mask RCNN are the most used detection algorithms in the conventional methods. SSD performs well for detecting large objects but performs worse for detecting small objects [7], which results in losing the small persons or the persons at a small spatial scale. Also, as mentioned previously, SSD suffers from false positive detections that confuse the person attribute recognition algorithm. Furthermore, mask RCNN is more accurate, but it is slower than SSD, and many authors used it in the conventional person retrieval techniques. However, the mask RCNN does not always succeed in predicting the instances details and producing an accurate object mask because it considers a portion of the background as foreground, which results in extreme boundaries segmentation [8,9]. Moreover, the mask RCNN cannot successfully detect all the objects in surveillance application due to crowd, occlusion, and variations in orientation, illumination, object scale, and object postures [10]. These reasons result in false detection, missed detection, or multiple bounding boxes for the same person. All these drawbacks affect the performance of the person attribute recognition and the retrieval systems.

Schumann et al. [2] exploit the movement direction and the speed information using a linear motion tracking model to reduce the SSD detector errors. However, the performance still needs to be improved because the tracking algorithm used in his method did not overcome all drawbacks.

The other factor that affects the person retrieval system is the person attribute recognition algorithm; if the person's attributes are not recognized correctly, the person will not be retrieved because of the unmatched query. The performance of attribute recognition algorithms is affected by viewpoint change, weather variations, low illumination, etc. Some methods, as mentioned earlier, use a single classifier for each attribute to guarantee better attribute recognition accuracy, while the other methods use one classifier for all attributes. A single classifier for each attribute is used in cascading order to get the final person attributes; however, its main problem is that if an error occurs in a classifier, the error will disseminate to the following classifiers. Further, combining different attributes in one classifier reduces attribute recognition accuracy due to attribute class imbalances [2].

In this paper, the proposed algorithm overcomes the detectors' faults by exploiting the spatio-temporal relation of the person bounding boxes to link them through each sequence using the ByteTrack algorithm [11]. First, the ByteTrack applies the YOLOX detector [12] on each sequence to detect the persons. Then, ByteTrack generates a tube for each person by associating the bounding boxes via the Byte algorithm. The advantage of the Byte is that it does not only link the high confidence score boxes to the tracklets but also links the low score boxes to the unmatched tracklets to filter out the correct objects from the false positive detections. Due to ByteTrack's precise detection and low confidence score boxes association, the ByteTrack is more robust to occlusions, reduces the identity switch, and reserves the identities successfully. In addition to reducing the detectors' drawbacks, separating the persons in tubes provides rich information about each person and improves attribute recognition. In case of the attribute recognition algorithm fails to recognize appropriately-one or more attributes for one of the bounding boxes of a person because of occlusions, the proposed algorithm uses the attribute value of the other bounding boxes in the same tube to decide the value of this attribute and this help in retrieving the right person. Additionally, the proposed algorithm increases the accuracy of attribute recognition by merging two attribute recognition algorithms to exploit their advantages and select the correct attributes from both. The two algorithms are Attribute-person recognition (APR) [13] which is a global image-based technique that extracts features from the whole image that describe the texture, color, and shape, while

Attribute Localization Module (ALM) [14] is a part-based technique that acquires the regional features for each attribute.

The main contributions of this proposed research are summarized as follow:

1. A person retrieval algorithm that incorporates the ByteTrack algorithm for object detection and APR & ALM for person attribute recognition to improve the retrieval accuracy in surveillance applications is proposed.
2. A ByteTrack algorithm is utilized instead of the object detection methods to achieve two preferences:
   - Overcome the false positive and missing detections by applying a precise and fast detector YOLOX and accurate Byte association algorithm to the video sequences.
   - Exploit the spatio-temporal of the person to improve the attributes fault results from the attribute recognition algorithms.
3. The predictions of APR and ALM person attribute recognition algorithms, trained on two different pedestrian datasets, are merged to fulfill a better accurate attribute classification result.
4. Extensive experiments accomplished on the SoftBioSearch dataset have revealed the superior performance of the presented approach.

The rest of the paper is structured as follows. Section 2 presents backgrounds about the main modules of the person retrieval algorithms. Section 3 describes the proposed person retrieval algorithm and its modules. The experimental and simulation results are discussed in section 4. In the end, Section 5 concludes the proposed research work.

2. Background

The person retrieval framework based on semantic descriptions incorporates several modules, including object detection, object tracking, and attribute recognition. This section describes the algorithms that can be used in each one of these modules.

2.1. Object detection

Object detection task endeavors to locate perceptible object instances in images or videos. Commonly, each detected object is bordered by a bounding box tagged with a class label and a corresponding confidence value. Person detection is an alternative form of object detection utilized to capture the principal persons in images or video frames. Detecting individuals in videos is deemed a crucial task in recent video surveillance frameworks. The contemporary deep learning techniques proffer robust approaches for person detection [15]. Generally, object detection techniques fall into two principal strategies: One-stage approaches, which utilize a single network to fulfill the object localization and classification, and two-stage approaches, which enclose two detached networks to achieve the localization and classification tasks [16]. The first category of approaches includes some popular detectors, such as YOLO (You Only Look Once) series [17], and SSD [18].

YOLO is an object detection technique that employs a CNN single-stage architecture revealing multiple objects in real-time. The CNN is applied to predict the confidence values and bounding boxes assigned to each detected object. There are various variants of YOLO algorithms. Redmon et al. were the first to propose YOLO v1 for object detection [19]. Afterward, they benefited from the most evolved detection technologies available to improve the overall performance. YOLO v2 utilized anchor boxes to capture the objects in an image [20], while YOLO v3 implemented Residual Net [21]. YOLO v4 was released by Bochkovskiy et al. [22] to improve YOLO v3. YOLO v5 [17] was introduced pretty soon after YOLO v4. This achievement has a similar design to YOLO v4, yet, YOLO v5 models proved to be faster to train, performant, and user friendly. Currently, Ge et al. released YOLOX, which fulfills a superior balance between speed and accuracy compared to their counterparts [12]. Similarly, SSD incorporates procedures for locating and classifying bounding boxes of each detected object [18]. SSD employs the VGG-16 architecture as a base net to extract features from the input image. Besides, several convolutional filters are subsequently combined to extract multi-scale feature maps for augmenting the detection speed while maintaining accuracy.

Among methods of the second approach are Region-based Convolutional Neural Networks (R-CNN) [18], Fast R-CNN [23], Faster R-CNN [24] and Mask RCNN [25]) algorithms. R-CNN is a hybrid technique that applies both traditional methods and learning-based approaches. It employs selective search to capture information regarding the region of interest (ROI). For each region, CNN is used to extract the enclosed features. These features are classified using the support vector machine (SVM) classifier to classify the objects within the ROIs. Subsequently, Girshick et al. submitted Fast R-CNN [23], a modified version of R-CNN. This technique proposed ROI pooling layers, which decreases the computational cost by using a single feature map for all regions. Furthermore, it also applied a deep learning method for classification and a regression network for refining the bounding boxes. Thereafter, Ren et al. raised Faster R-CNN [24], in which the selective search technique is replaced with an additional CNN to achieve the regional proposal. This yields a much faster algorithm. The acquired regional proposals are fed then to the ROI pooling module for further procedures. He et al. [25] fulfilled Mask R-CNN as an improved version of Faster R-CNN and acted by performing parallel attachment between a branch for anticipating an object mask and the existing bounding box recognition branch. This idea increases the algorithm's accuracy and reduces the processing time. Generally, one-stage techniques are faster than two-stage techniques with comparable performance [26]. Thus, in this paper, the YOLOX algorithm has been applied for person detection to attain better competence.

2.2. Object tracking

Object tracking can be categorized into; single object tracking and multiple object tracking (MOT). MOT is a vigorous and challenging research subject, and it aims to speculate on the trajectories of the objects presented in videos. Diverse approaches were introduced to tackle the MOT problem.

During past years, many published studies adopted Correlation Filters (CFs) for visual tracking due to their effective computation capability [27–30]. However, these tracking

techniques utilize a single correlation filter to follow the target, which restricts the accuracy and robustness of the tracking procedure. Zhang et al. employed cascaded CFs to enhance tracking accuracy [31].

Furthermore, some researchers focused on using handcraft features such as Color Names (CN) [32], Scale-Invariant Feature Transform (SIFT) [33] and Histogram Oriented Gradient (HOG) [34]. In the work of [35], color histogram and HOG are fused to achieve complementary tracking. The handcrafted features are typically not rugged, sensitive to environmental conditions, and have high computational costs due to high dimensions.

Recently, deep learning techniques embraced achieving object tracking. Modern studies verified that trackers based on deep learning are efficacious in enhancing the tracking performance regarding the tracking prediction and data association aspects. The commonly MOT strategy utilized is tracking-by-detection, in which the tracking task is divided into two stages, a detection stage and an association stage. In the first stage, the bounding boxes, which specify the targets in the video, are extracted. These bounding boxes are employed to drive the tracking process. The association stage assigns the same ID for bounding boxes that comprise the same target [36]. Lately, a thoroughly different procedure based on Siamese networks has been presented [37–39]. This technique is relied on training a similarity function offline on a couple of video frames instead of performing online learning for the discriminative classifier. Furthermore, Shuai et al. [40] blend the Faster-RCNN as a region-based detection method with a Siamese-based object tracker. The Siamese network is employed to model the motion instance across frames.

Some works incorporate the deep learning techniques and CFs to improve the tracking performance. Zhang et al. [31] submit a tracking algorithm based on extracting features utilizing ResNet features and combining them with cascaded CFs. Others adopted combining deep and handcrafted features. Yang et al. [41] propose an object tracking approach based on hierarchical features comprised of handcrafted features: HOG and CN, and deep features acquired from three layers of pre-trained VGG19.

Wojke et al. [42] offer the DeepSORT framework as an extension to Simple Online and Realtime Tracking (SORT) technique. They incorporate motion and appearance information to enhance the performance of SORT. This technique became popular and widely used [43]. Most of the proposed MOT techniques determine the identity of objects by associating bounding boxes possessing scores exceeding a specific threshold. This has led to the disappearance of certain objects whose detection scores are low. Thus, to resolve this issue, Zhang et al suggested an MOT technique based on a simple, efficient, and inclusive association procedure, where a high tolerance approach is used when selecting detection results at the association stage [11]. This algorithm has rendered a high-performing association, hence its adoption in this work.

2.3. Attribute recognition

Classical techniques of recognizing attributes of the person usually focus on elaborating a powerful representation of the characteristics from the viewpoints of the handcrafted characteristics, robust classifier systems, or attribute relationships. However, the evaluations on large-scale benchmarks imply that the effectiveness of these conventional approaches is far from the realistic application requirements [44]. Over the last few years, deep learning has attained exceptional performance due to its successful ability for the automatic extraction of features. Most of the existing attribute recognition methods can be categorized into global image-based, part-based and attention-based models. In global image-based models [45–48], the algorithms are applied to the whole images and do Person attribute recognition multi-task learning. The advantages of these models are as follows: simple, intuitive, and extremely efficient; however, these methods performance is not accurate due to ignoring the fine-grained features. In the part-based models [14,49–51], the techniques localize the body parts based on part detection or proposal region generation methods to get the fine-grained local features. Concentrating on the person parts improves the person attribute classification performance however; these methods performance is affected by the accuracy of the part detection algorithms. In the attention-based models [52–55], the methods use the attention idea which is the behavioral and cognitive act of focusing on a specified component of information, whether subjective or objective, while disregarding other perceptible information. In this paper, the APR algorithm from the global-based models and the ALM algorithm from the part-based models are merged to exploit their advantages to improve the attribute recognition.

3. Methodology

This section proposes an approach for a spatio-temporal person retrieval algorithm based on a witness description. Most of the aforementioned conventional retrieval methods use a detector, such as Mask-RCNN and SSD, as the first module in the retrieval algorithm; and a person attribute recognition as a second module to recognize the person's attributes. However, these methods suffer from some detection faults such as false positives, missed detections, and generating multi bounding boxes for the same person, which affect the retrieval performance. Furthermore, Defects in the detectors increase with insufficient illumination, occlusion, and variation in the object scale, especially for surveillance applications. The proposed algorithm overcomes these drawbacks by taking advantage of the temporal relations between the persons bounding boxes and creating a tube for each person. The generated tube carries more information about the person that helps retrieve him effectively better than using each bounding box alone. Classification of attributes for each bounding box alone for the same person may vary due to illumination, occlusion, and object scaling, resulting in missing one or more bounding boxes. However, for attribute classification based on the tube, the final attribute value for that person depends on the attribute values of all the bounding boxes in the tube, which handles shortcomings in attributes recognition.

On the other hand, most of the conventional attribute recognition algorithms used in the second module cannot classify all pedestrian attributes correctly due to the surveillance challenges that affect the person retrieval. The proposed method exploits the advantages of two attribute recognition algorithms and merges their predictions. The proposed retrieval algorithm, the tracker module, and the person attribute

recognition module will be described in detail in the following subsections.

3.1. The proposed person retrieval algorithm

Fig. 1 demonstrates the complete structure of the proposed retrieval algorithm and its modules. The suggested method is designed based on two main modules: the tracker and the person attribute recognition. A ByteTrack algorithm is exploited in the tracker module, while APR and ALM are utilized in the attribute recognition module. The query attributes are then matched to the recognized attributes of each tube to retrieve the required person. The retrieval algorithm steps are as follows:

Step 1: Apply the Bytetrack algorithm to extract the person's tubes on each surveillance video.

Step 2: Apply the APR and ALM on each bounding box of the person in the tube to recognize their attributes.

Step 3: Calculate the final attribute value for each person depending on all the bounding boxes' attribute values in the tube according to the following equation:

$$a_r = \frac{\sum_{i=1}^{k} a_r^{f_i}}{k} \tag{1}$$

Where a_r is the average occurrence of a single attribute all over the tube for a person, k is the number of frames within the tube, $a_r^{f_k}$ is the predicted attribute from n attributes for a person at the frame f_k which take the following values:

$$a_r^{f_k} = \begin{cases} 1 & \text{if predict attribute is recognized in frame} f_k \\ 0 & \text{if predict attribute is unrecognized in frame} f_k \end{cases} \tag{2}$$

Step 4: Calculate the distance between the query attributes and the recognized attributes and retrieve the person tube with minimum distance.

3.2. ByteTrack module

The goal of MOT is to estimate object bounding boxes and identities in videos. Most approaches work by associating detection boxes with scores above a certain threshold; however, Low-detection-score objects, such as occluded objects, are simply discarded, leaving significant genuine objects missing and broken trajectories. ByteTrack addresses this issue by proposing a simple, effective, and flexible association mechanism that tracks practically every detection box rather than only the high-scoring ones. Then, it uses the similarities of the low score detection boxes with tracklets to retrieve actual objects and filter out background detections. This makes the ByteTrack robust to object occlusion, reduces the identity switch, and saves the identities. ByteTrack uses the high-performance detector YOLOX to conduct MOT on a video and BYTE to associate the detection boxes and the tracks with creating tubes. The following subsections illustrate the process of the YOLOX and Byte association algorithm.

- Yolox

YOLOX is an improved version of YOLO [19] in a simple scheme and better performance without the anchor mechanism. However, the anchor mechanism in YOLO affects its model speed and computational cost. Also, the classification and regression tasks in YOLO (bounding box location estimation) are carried out simultaneously, which is known to generate conflicts and lower accuracy. YOLOX overcomes these issues by eliminating box anchors, resulting in lower computing costs and faster inference. Furthermore, YOLOX decouples the YOLO detection head into distinct feature channels for box coordinate regression and object classification. As a result, model accuracy and convergence speed are improved.

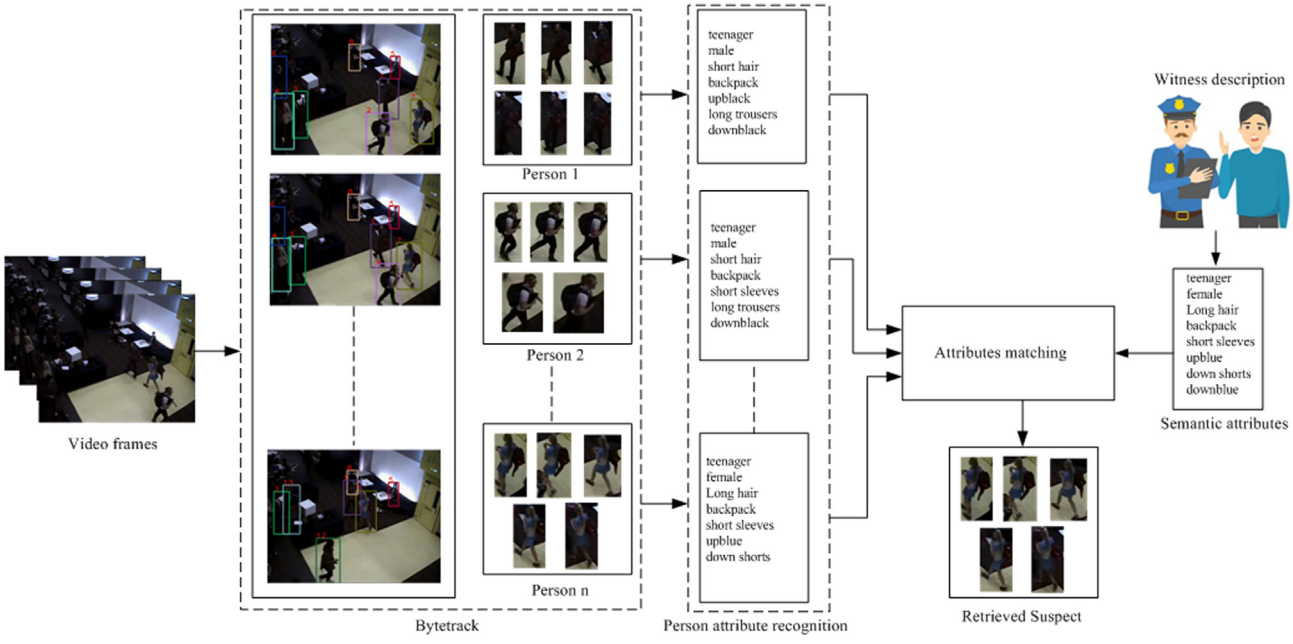


Fig. 1 The proposed spatio-temporal person retrieval based on witness description.

• Byte association method

Most methods of MOT estimate the identities of objects in videos by relating the detected bounding boxes whose confidence scores are greater than a threshold. However, the bounding boxes with low confidence scores which are produced due to occlusion, blurring, lighting effect, or scaling the object, are ignored even though the object may be present, leading to disjointed trajectories. On the other hand, associating these low score bounding boxes adds true positives and produces false positives. Byte solves this issue by first linking the high score boxes to the tracklets and then linking between unmatched tracklets and the low score boxes to filter out the correct objects from the false positive detections. First, the BYTE uses the Kalman filter to predict in the next frame the position of the objects in the tracklets, and then, two matching processes are applied to associate the bounding boxes. The first matching process uses the intersection over union (IoU) to match the predicted positions with the high confidence score bounding boxes. The second matching process matches the unmatched objects in the tracklets and the low score bounding boxes. This causes the low score detection boxes to be properly matched to the preceding tracklets, and the objects are restored. At the same time, the false positive (background) is neglected, as shown in Fig. 2 [11].

3.3. Person attribute recognition module

Recognizing the visual attributes of the pedestrian from surveillance video is a challenging task due to insufficient illumination, person pose variations, weather condition, and person occlusion. The proposed method uses the benefits of two attribute recognition algorithms and combines their predictions to improve the attribute prediction. One of these algorithms is a global image-based technique called APR that takes the whole image as an input and extracts the features that describe the overall properties of this image, such as texture, color, and shape. The other algorithm is part-based, called ALM that learns the regional features for each attribute at several levels and adaptively discovers the best discriminative locations. The proposed algorithm selects empirically the attributes that are accurate from the ARP as the.

color and merges them with the attributes that are accurate from ALM as the type of clothes and hair length to guarantee better retrieval performance. Eventually, the predicted attributes for each person are matched to the attribute query to retrieve the person based on this query. In the following subsection, APR and ALM algorithms are described.

3.3.1. Apr

In this algorithm, CNN is used to extract the pedestrian features, and these features are then fed to a classifier to predict the attributes. The algorithm uses ResNet-50 as a backbone network and is trained on Market-1501, one of the large-scale surveillance datasets [56].

3.3.2. Alm

The algorithm consists of the main network with feature pyramid structures and a collection of Attribute Localization Modules based on the Spatial Transformer Network (STN) [57] applied to different feature levels. The algorithm extracts the features of each pedestrian bounding box using the Batch Normalization inception (BN-inception) network as a backbone network and feds the collected features from different levels into several ALM to obtain the attribute vector. Each ALM performs attribute localization and region-based feature learning for only one attribute at a single level. The algorithm is trained on the most familiar pedestrian attribute dataset PETA [58].

4. Experiments

This section evaluates the proposed method qualitatively and quantitatively compared to the conventional methods by conducting different experiments on a SoftBioSearch database. The dataset is composed of several videos, and each video implies retrieving a particular person. The person retrieval system accuracy is measured by how precise this person is located all over the video. The detailed information on the evaluation metrics, the dataset, and the simulation results are covered in the following subsections.

4.1. Evaluation metric

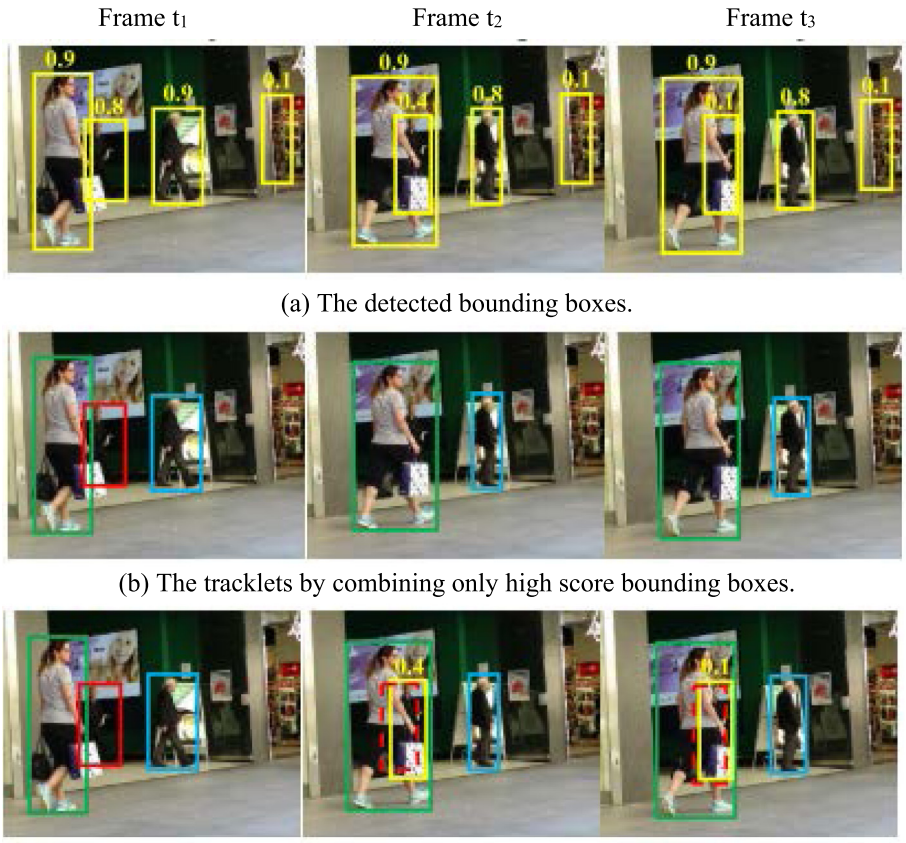
The metrics used to evaluate the proposed method are the average IoU and percent of frames with an IoU ≥ 0.4, which is the percentage of true positive retrieval "TP." IoU is used because the person retrieval methods performance depends on the accuracy of the person localization of the detector. IoU is calculated by dividing the intersection area of the ground-truth bounding boxes and the predicted bounding boxes by their union area, as shown in Fig. 3. First, the IoU is calculated for the retrieved person per video sequence; then, a final IoU is obtained by averaging the video sequences results by the number of video sequences.

4.2. Dataset

The proposed approach is evaluated on the SoftBioSearch database [59], which is captured based on six cameras and consists of 110 video sequences for training and 41 video sequences for testing. These six cameras are calibrated using Tsai's camera calibration technique [60]. Each video sequence concentrates only on a single object surrounded by a bounding box as shown in Fig. 4, and has a semantic query that defines this object (gender, age, height, build, hair and skin color, clothing type, texture, and color). Extensive diversity of objects is collected with different colors, heights, and appearances by recording the video at six distinct periods and different crowding levels. The recorded videos are selected on a scale from easy to difficult. Easy videos contain clear and visible persons, medium videos have two or more persons but are still distinguishable, and difficult videos have a congested scene with heavy occluded persons. Image samples from the recorded video from each of the six cameras are shown in Fig. 4 [59].

4.3. Simulation results

The proposed algorithm localizes and retrieves persons in surveillance videos using a semantic query. According to the query, the person is located in all the frames where the person appears. Following the conventional person retrieval method,



(a) The detected bounding boxes.

(b) The tracklets by combining only high score bounding boxes.

(c)The tracklets by combining each bounding box.

Fig. 2 Examples of ByteTrack method. (a) The detection boxes with their scores. (b) Conventional methods tracklets which are obtained by associating high score detection boxes. (c) ByteTrack tracklets. The dashed boxes reflect the Kalman Filter predicted bounding box of the previous tracklets. The previous tracklets are appropriately matched to the two low-score detection boxes.

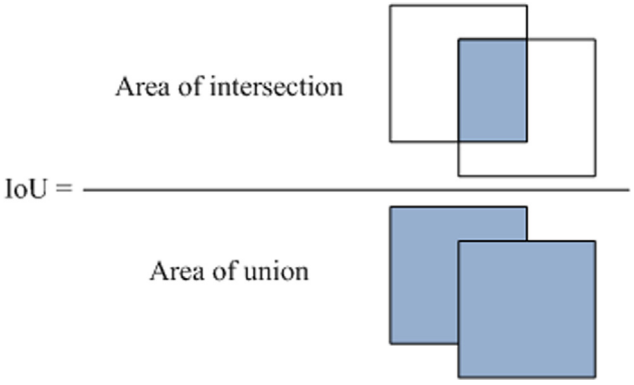


Fig. 3 IoU metric.

the test set consisting of 41 persons is used to evaluate the proposed method compared to the conventional methods. According to [4], these 41 video sequences are divided into four levels of difficulty: 6 very easy, 13 easy, 12 medium, and 10 hard. Different descriptors are selected empirically to describe each person as age, gender, hair length, luggage, sleeves length, torso color and legs, clothes type, color, and length, which can retrieve the correct person. Additional descriptors are added as the type of clothes, formal or casual, sunglasses, and footwear, but it is noticed that these descriptors do not affect the retrieval accuracy. Fig. 5 shows the true positive retrieved persons based on a specific query. Three video sequences (sequences 0, 11, and 36) with different difficulty levels are chosen to present the proposed algorithm performance. The figure shows for each selected sequence: the query, the detected persons utiliz-

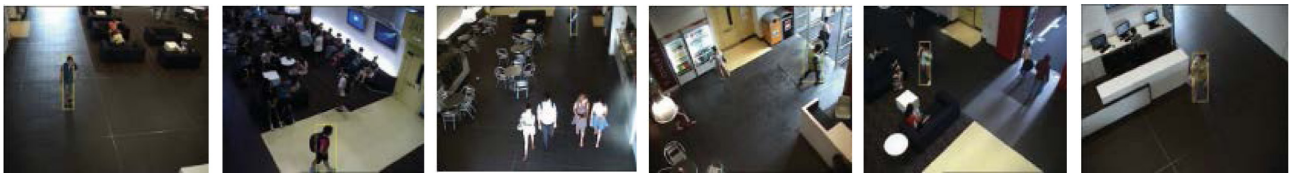


Fig. 4 An image sample from the video sequences for each camera.

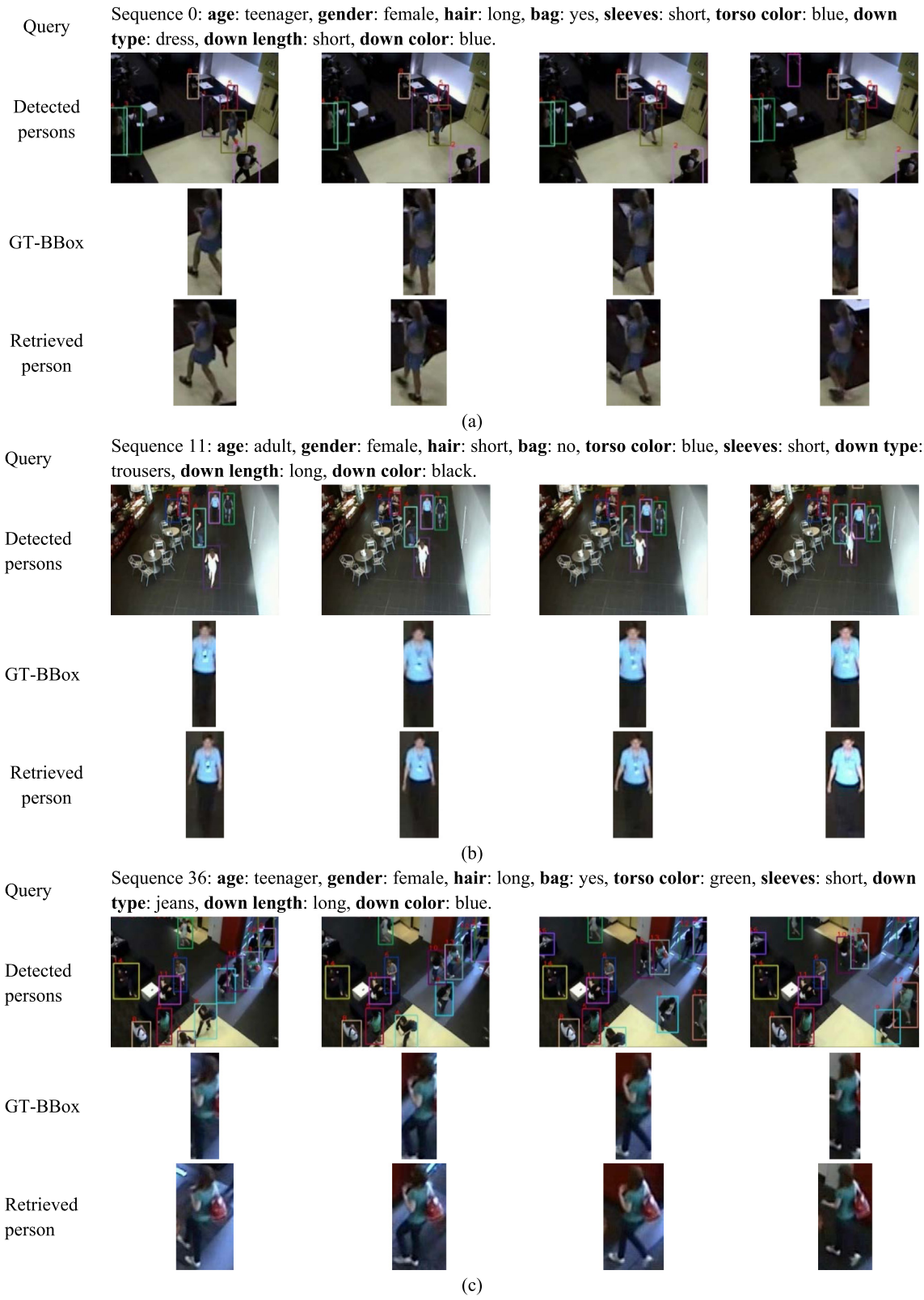


Fig. 5 Visualization of the proposed person retrieval performance on three different video sequences (a) video sequence 0 (b) video sequence 11 (c) video sequence 36.

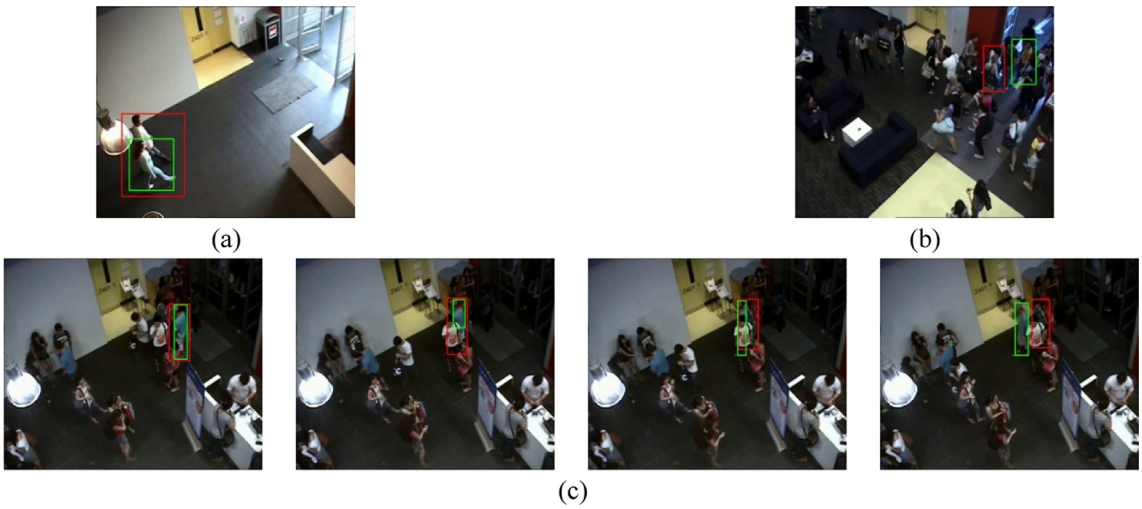


Fig. 6 The proposed personal retrieval method false negatives samples. Green: ground truth, red: retrieved person.

Table 1 The comparison of the proposed method and the conventional methods in terms of the modules, number of attributes and performance.

Approach	Module 1		Module 2	No of Attributes	Performance	
	Detection	Tracking	Attribute Recognition		Average IoU	IoU > 0.4 (TP)
Galiyawala et al. [1]	Mask R-CNN	–	Cascading linear filters	3	0.363	0.522
Schumann et al. [2]	SSD	Simple linear motion model	Ensemble of classifiers	9	0.503	0.759
Yaguchi et al. [3]	Mask R-CNN	–	Ensemble of classifiers	9	0.511	0.669
Galiyawala et al. [6]	Mask R-CNN	–	Cascading linear filters	4	0.569	0.746
Parshwa Shah et al. [4]	Mask R-CNN	–	Cascading linear filters	7	0.566	0.792
The proposed approach	YOLOX	ByteTrack	ALM + APR	8	0.561	0.9321

ing the ByteTrack algorithm, the ground-truth bounding box, and the retrieved person bounding box.

As earlier mentioned, most of the retrieval algorithms face many challenges due to bad illumination, occlusion, and heavily crowded area. ByteTrack overcomes these challenges but fails in a few frames and produces false-negative cases, which affect the retrieval system accuracy, as shown in Fig. 6. For example, it can be seen from Fig. 6 (a) that the proposed method retrieves two persons instead of the required person due to their complete occlusion. Also, the proposed algorithm fails to recognize the persons with approximately similar appearances due to the bad illumination and heavy crowd as in Fig. 6(b). On the other hand, in Fig. 6 (c), the smooth transition between the persons leads the detection algorithm to switch the person's identity for a few frames.

The proposed methodology is benchmarked against five traditional techniques [1], [2], [3], [6], [4] to assess its performance. The findings are set out in Table 1. It can be noticed from the table that most of the conventional methods use detection algorithms and only one method uses a tracking algorithm. Also, it can be observed from the results that the average IoU of the proposed method exceeds [1], [2], [3] and is approximately equal to [6] and [4]. On the other hand, the proposed method outperforms all the conventional methods in terms of TP. However, the results raise a question: "why is IoU a little bit small while TP is high". Fig. 5 answers this question by observing that the ByteTrack output is more accurate than the ground-truth, since the bounding box surrounds the whole person as well as his attachments (e.g. holding a bag). However, the ground-truth bounding box is very tight, and parts of the persons are cropped. This results in decreasing the IoU for the proposed method. On the other hand, it can be remarked from the table that TP is higher than that of the best conventional technique by 14 %. Fig. 7 presents the TP percentage of the proposed technique for each sequence. The proposed system achieved TP = 100 % in 23 of 41 video sequences and TP between 70 % and 100 % for the rest of the 18 sequences. The results reflect that regardless of whether the video sequences are very crowded and severe from high occlusion, the proposed method successfully retrieves the person from most of the frames in the video sequences.

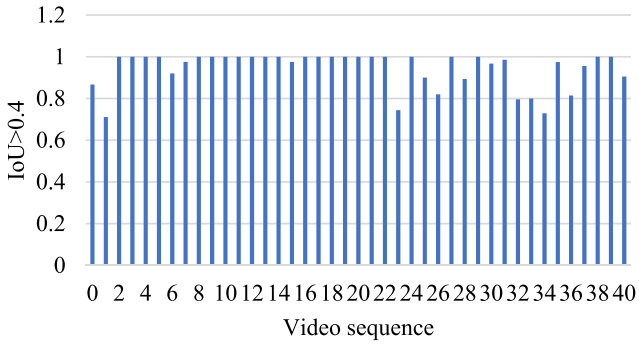
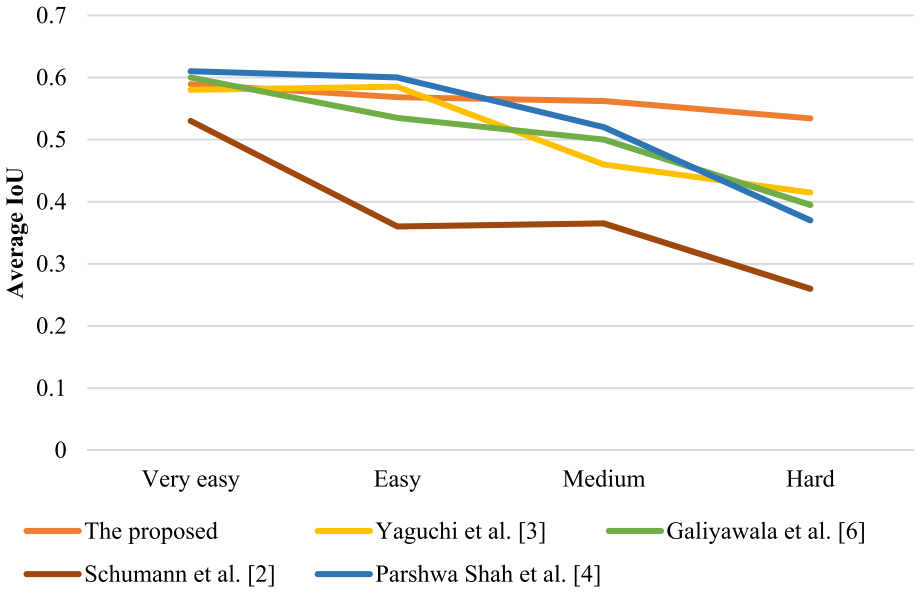


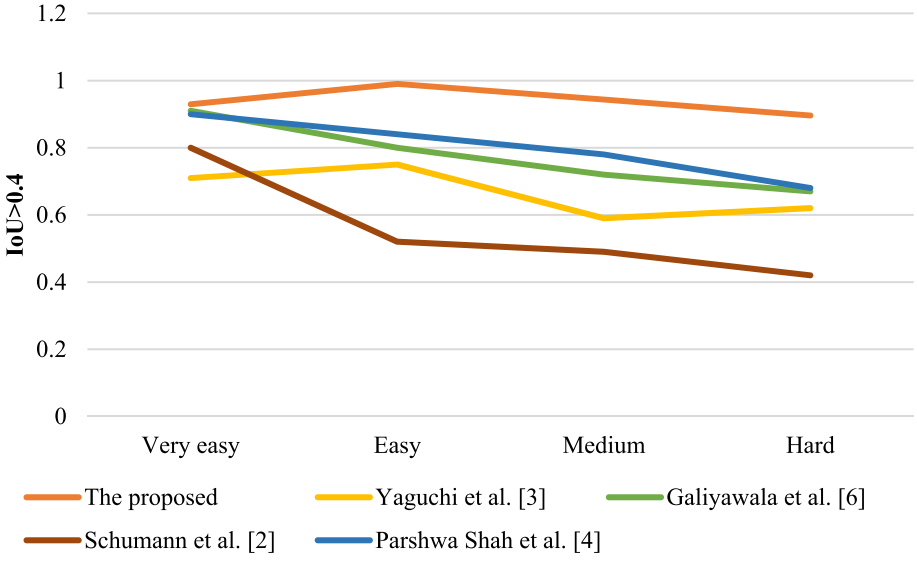
Fig. 7 The performance of the proposed algorithm for each video sequence in terms of TP.

Moreover, the effectiveness of the proposed method on different levels of video difficulty compared to the conventional methods in terms of average IoU and TP is shown in Fig. 8. It can be perceived from the figure that the proposed method performs better than the other conventional methods as the difficulty of the videos increase and its results are comparable for very easy and easy video sequences.

As mentioned previously, the tracking stage is added to the proposed retrieval system to improve its performance because the occluded persons lose their body parts and affect the person attributes recognition. Thus, the query does not match the person's attributes and will not be retrieved. The role of the tracking stage is to link all the bounding boxes of each per-



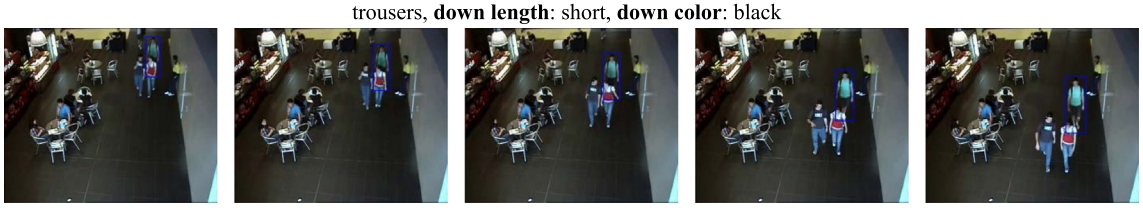
(a)



(b)

Fig. 8 The performance of the proposed technique compared to the conventional methods with respect to the videos difficulty levels in terms of (a)average IoU (b) IoU > 0.4.

Query: **age**: teenager, **gender**: male, **hair length**: short, **luggage**: yes, **sleeves length**: short, **torso color**: green, **down type**: trousers, **down length**: short, **down color**: black



(a)	(b)	(c)	(d)	(e)
Age: teenager	**Age**: teenager	**Age**: teenager	**Age**: teenager	**Age**: teenager
Gender: male	**Gender**: male	**Gender**: male	**Gender**: male	**Gender**: male
Hair length: short	**Hair length**: short	**Hair length**: short	**Hair length**: short	**Hair length**: short
Luggage: yes	**Luggage**: yes	**Luggage**: yes	**Luggage**: yes	**Luggage**: yes
Sleeves length: short	**Sleeves length**: short	**Sleeves length**: short	**Sleeves length**: short	**Sleeves length**: short
Torso color: green	**Torso color**: green	**Torso color**: green	**Torso color**: green	**Torso color**: green
Down type:--	**Down type**:--	**Down type**: trousers	**Down type**: trousers	**Down type**: trousers
Down length: --	**Down length**: --	**Down length**: --	**Down length**: short	**Down length**: short
Down color: --	**Down color**: --	**Down color**: --	**Down color**: black	**Down color**: black

Fig. 9 The effect of the tracking stage on the performance of the proposed person retrieval algorithm. The attributes of: (a) frame no. 20 (b) frame no. 40 (c) frame no. 60 (d) frame no. 90 (e) frame no. 100.

Table 2 The effect of the attribute recognition algorithms on the performance of the proposed method.

Approach	Average IoU	IoU > 0.4 (TP)
ALM	0.4060	0.6727
APR	0.5279	0.8807
APR + ALM (proposed)	0.561	0.9321

son together. Hence, when a group of the bounding boxes matches the query, this person's occurrences are retrieved, as shown in Fig. 9. However, it is noticeable from the figure that frames numbers: 20, 40 and 60 miss attributes for the occluded parts, so the person will not be retrieved according to the query. Nevertheless, the tracking module relates the person bounding box to solve the occluded part issue.

4.4. Ablation study

4.4.1. The effect of the attribute recognition algorithms

This section analyzes the proposed algorithm according to the influence of the attribute recognition algorithms. The proposed algorithm is implemented using ALM & APR, APR, and ALM. The proposed algorithm is applied to the 41 video sequences and evaluated using the IoU metric. Table 2 presents the results that reflect the performance of the proposed method. It can be remarked from the results that merging the predictions from APR and ALM improves the retrieval performance in terms of IoU and TP compared to using only one of them. Fig. 10 confirms this observation. It can be noticed from the figure that for the sequences 5, 8, 21, 24, 25, 27, 29, 30, 32, 33, 36, 37, and 41, sometimes APR and ALM fail to retrieve the required person. However, merging their predictions strengthens the proposed algorithm and helps it retrieve the persons in all the video sequences.

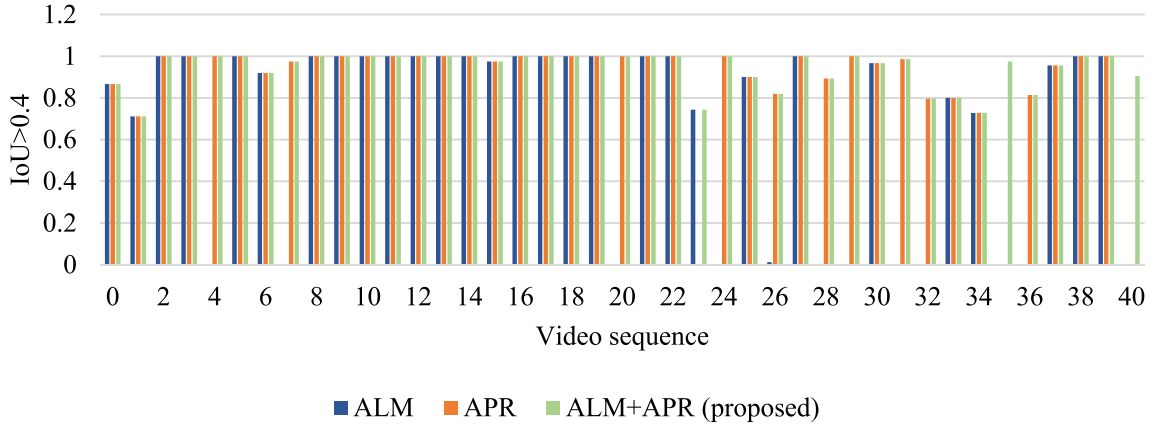


Fig. 10 The performance of the proposed algorithm using different attribute recognition algorithms in terms of IoU > 0.4.

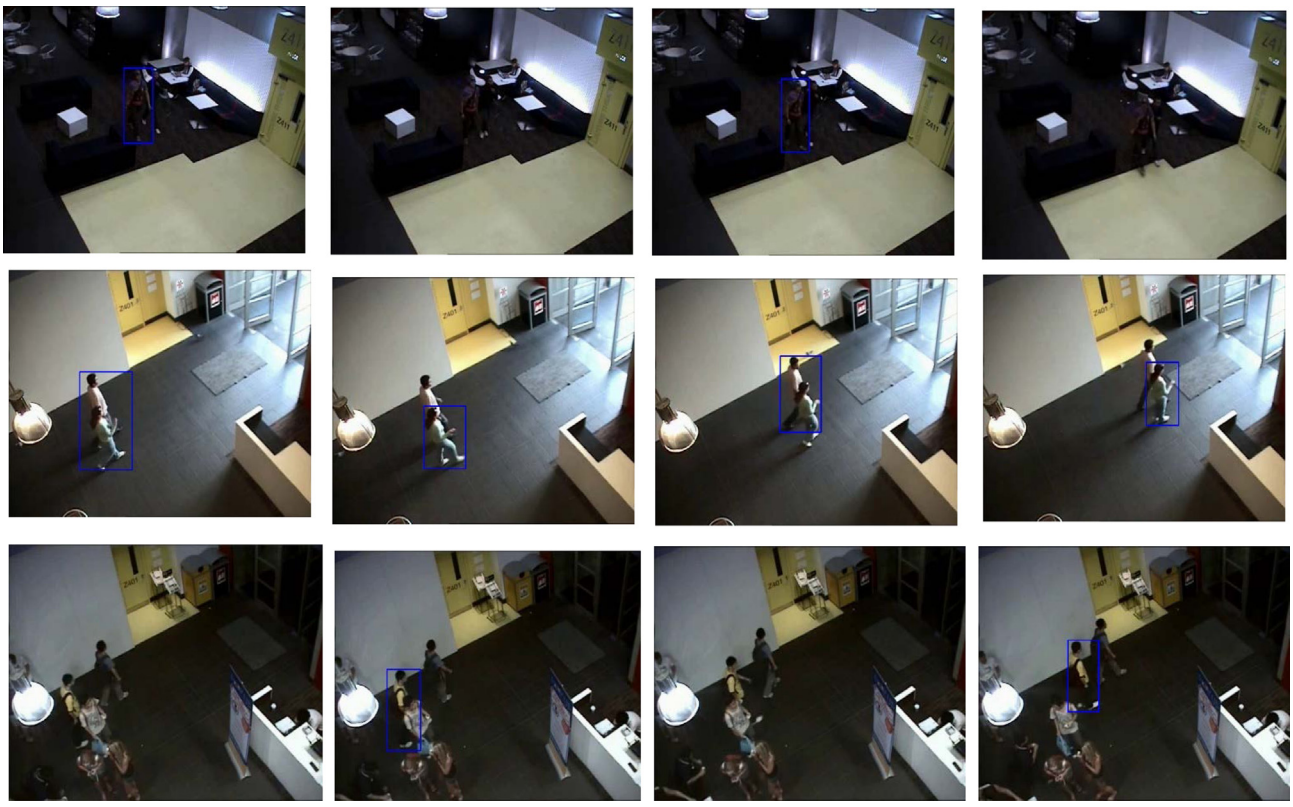


Fig. 11 The effect of neglecting low confidence score bounding boxes in the proposed method in case of bad illumination (first row), occlusion (second row) and crowded area (third row).

Table 3 The performance of the proposed method at different IoU thresholds.

	The proposed algorithm
IoU > 0.3	0.99
IoU > 0.35	0.97
IoU > 0.4	0.93
IoU > 0.45	0.88
IoU > 0.5	0.76
IoU > 0.55	0.57
IoU > 0.6	0.35

4.4.2. The effect of associating low and high confidence score bounding boxes

Several experiments are conducted by neglecting the low confidence score bounding boxes to prove the effect of associating low and high confidence score bounding boxes in the tracking stage on the performance of the proposed retrieval algorithm, and the results are presented in Fig. 11. It can be shown from the figure that some bounding boxes are lost in case of bad illumination, occlusion, or crowded areas due to ignoring the low score bounding boxes. This decreases the average IoU value by 6 % to 39 % and decreases the TP value by 6 % to 70 % for complex videos.

4.4.3. The effect of the IoU threshold:

As previously described, the performance of the proposed method is measured by the percent of frames with an IoU $\geq$ 0.4 which is estimated by the conventional researches. The calculated value reflects the percentage of true positive retrieval where the proposed method succeeds. In this experiment, the proposed algorithm performance is evaluated at different thresholds for IoU. Table 3 presents the performance of the proposed method at different threshold values. The obtained results indicate that using IOU threshold = 0.4 is a reasonable value. Furthermore, this metric was devoted to the other person retrieval researches. Thus, in this work, we adopted this metric to compare the performance of the proposed technique with the other state-of-the-art methods.

5. Conclusion

This study suggested a person retrieval approach based on the ByteTrack strategy for object tracking. The proposed procedure aims to extract a tube for each person in the video frames. Each tube maintains the spatio-temporal information of its affiliated person. Extricating attributes from these tubes leads to more precise attributes recognition than depicting the attributes from each bounding box. Eventually, the person retrieval is performed by fusing two attribute recognition algorithms to enhance the recognition performance. A comprehensive empirical appraisal has been fulfilled, indicating the efficacy of the proposed algorithm. Furthermore, the proposed algorithm is compared to five traditional methods. The obtained results demonstrate that the offered technique outperforms these techniques due to its ability to overcome the occlusion and missing detection problems.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] H. Galiyawala, K. Shah, V. Gajjar, M.S. Raval, Person Retrieval in Surveillance Video using Height, Color and Gender, in: Proc. AVSS 2018 - 2018 15th IEEE Int. Conf. Adv. Video Signal-Based Surveill., 2019.

[2] A. Schumann, A. Specker, J. Beyerer, Attribute-based Person Retrieval and Search in Video Sequences, in: Proc. AVSS 2018 - 2018 15th IEEE Int. Conf. Adv. Video Signal-Based Surveill., 2019.

[3] T. Yaguchi, M.S. Nixon, Transfer Learning Based Approach for Semantic Person Retrieval, in: Proc. AVSS 2018 - 2018 15th IEEE Int. Conf. Adv. Video Signal-Based Surveill., pp. 1–6, 2019.

[4] P. Shah, A. Garg, V. Gajjar, PeR-ViS: Person Retrieval in Video Surveillance using Semantic Description, in: Proc. - 2021 IEEE Winter Conf. Appl. Comput. Vis. Work. WACVW 2021, pp. 41–50, 2021.

[5] R.Y. Tsai, A Versatile Camera Calibration Techniaue for High-Accuracy 3D Machine Vision Metrology Using Off-the-shelf TV Cameras and Lenses, no. 4, 1987.

[6] H. Galiyawala, M.S. Raval, Person retrieval in surveillance using textual query: a review, Multimed. Tools Applications (2021).

[7] A. Kumar, Z.J. Zhang, H. Lyu, Object detection in real time based on improved single shot multi-box detector algorithm, EURASIP J. Wirel. Commun. Netw. 2020 (1) (2020), https://doi.org/10.1186/s13638-020-01826-x.

[8] G. Khan, Z. Tariq, J. Hussain, M.A. Farooq, M.U.G. Khan, Segmentation of crowd into multiple constituents using modified mask R-CNN based on mutual positioning of human, in: 2019 Int. Conf. Commun. Technol. ComTech 2019, no. ComTech, pp. 19–25, 2019.

[9] X. Wu, S.W.B, Y. Xie, Improvement of Mask-RCNN Object Segmentation Algorithm Improvement of Mask-RCNN Object, no. August. Springer International Publishing, 2019.

[10] U. Gawande, K. Hajari, Y. Golhar, SIRA: Scale illumination rotation affine invariant mask R-CNN for pedestrian detection, Appl. Intell. 0123456789 (2022).

[11] Y. Zhang et al, ByteTrack: Multi-Object Tracking by Associating Every Detection Box, arXiv Prepr. (2021).

[12] Z. Ge, S. Liu, F. Wang, Z. Li, J. Sun, YOLOX: Exceeding YOLO Series in 2021, arXiv:2107.08430, pp. 1–7, 2021.

[13] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, Improving person re-identification by attribute and identity learning, Pattern Recognit. J. 95 (2019) 151–161.

[14] C. Tang, L. Sheng, Z. X. Zhang, X. Hu, Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization, in: Proc. IEEE Int. Conf. Comput. Vis., vol. 2019-Octob, no. c, pp. 4996–5005, 2019.

[15] H.D. Najeeb, R.F. Ghani, A Survey on Object Detection and Tracking in Soccer Videos, vol. 8, no. 1. Springer Singapore, 2021.

[16] A. Mauri et al, Deep learning for real-time 3D multi-object detection, localisation, and tracking: application to smart mobility, Sensors (Switzerland) 20 (2) (2020) 1–15.

[17] Glenn Jocher, "yolo5." https://github.com/ultralytics/yolov5, 2021.

[18] L. Wei, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, F.u. Cheng-Yang, SSD: single shot multibox detector wei, Eccv 1 (2016) 398–413.

[19] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.

[20] J. Redmon, A. Farhadi, YOLO9000: Better, faster, stronger, in: Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, vol. 2017-Janua, pp. 6517–6525, 2017.

[21] J. Redmon, A. Farhadi, YOLOv3: An Incremental Improvement," arXiv Prepr. arXiv1804.02767, 2018.

[22] A. Bochkovskiy, C.-Y. Wang, H.-Y. M. Liao, YOLOv4: Optimal Speed and Accuracy of Object Detection, arXiv Prepr. arXiv2004.10934, 2020.

[23] R. Girshick, Fast R-CNN, in: Proc. IEEE Int. Conf. Comput. Vis., vol. 2015 Inter, pp. 1440–1448, 2015.

[24] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1137–1149.

[25] K. He, G. Gkioxari, P. Dollar, R. Girshick, Mask R-CNN, Proc. IEEE Int. Conf. Comput. Vis. 2017-Octob (2017) 2980–2988.

[26] S. Patel, A. Patel, Object detection with convolutional neural networks, Lect. Notes Networks Syst. 141 (2021) 529–539.

[27] M. Danelljan, G. Hager, F.S. Khan, M. Felsberg, Learning spatially regularized correlation filters for visual tracking, Proc. IEEE Int. Conf. Comput. Vis. 2015 Inter (2015) 4310–4318.

[28] Y. Li, J. Zhu, A scale adaptive kernel correlation filter tracker with feature integration, Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 8926 (2015) 254–265.

[29] S. Liu, D. Liu, G. Srivastava, D. Połap, M. Woźniak, Overview and methods of correlation filter algorithms in object tracking, Complex Intell. Syst. 7 (4) (2021) 1895–1917.

[30] F. Li, C. Tian, W. Zuo, L. Zhang, M.H. Yang, Learning spatial-temporal regularized correlation filters for visual tracking, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (2018) 4904–4913.

[31] J. Zhang, J. Sun, J. Wang, X.G. Yue, Visual object tracking based on residual network and cascaded correlation filters, J. Ambient Intell. Humaniz. Comput. 12 (8) (2021) 8427–8440.

[32] M. Danelljan, F.S. Khan, M. Felsberg, J. Van De Weijer, Adaptive color attributes for real-time visual tracking, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (2014) 1090–1097.

[33] Y. Wang, X. Luo, L. Ding, J. Wu, Object tracking via dense SIFT features and low-rank representation, Soft Comput. 23 (20) (2019) 10173–10186.

[34] R. Rai, S. Shukla, B. Singh, Histograms of Oriented Gradients for Human Detection, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005.

[35] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, P.H.S. Torr, Staple: Complementary learners for real-time tracking, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2016-Decem (2016) 1401–1409.

[36] Y. Xu, X. Zhou, S. Chen, F. Li, Deep learning for multiple object tracking: a survey, IET Comput. Vis. 13 (4) (2019) 411–419.

[37] L. Bertinetto, J. Valmadre, J.F. Henriques, A. Vedaldi, P.H.S. Torr, Fully-convolutional siamese networks for object tracking, Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) vol. 9914 LNCS (2016) 850–865.

[38] R. Tao, E. Gavves, A.W.M. Smeulders, Siamese instance search for tracking, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016, vol. 2016-Decem, pp. 1420–1429.

[39] A. He, C. L. B, X. Tian, W. Zeng, Towards a Better Match in Siamese Network Based Visual Object Tracker,in: European Conference on Computer Vision ECCV 2018, 2018, pp. 132–147.

[40] B. Shuai, A. Berneshawi, X. Li, D. Modolo, J. Tighe, Siammot: Siamese multi-object tracking, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (2021) 12367–12377.
[41] Y. Yang et al, Visual tracking with long-short term based correlation filter, IEEE Access 8 (2020) 20257–20269.
[42] N. Wojke, A. Bewley, D. Paulus, Simple online and realtime tracking with a deep association metric, Proc. - Int. Conf. Image Process. ICIP vol. 2017-Septe (2018) 3645–3649.
[43] S. K. Pal, A. Pramanik, JM. Pabitra, Deep learning in multi-object detection and tracking: state of the art, no. February. Applied Intelligence, 2021.
[44] X. Wang et al, Pedestrian attribute recognition: a survey, Pattern Recognit. 121 (2022) 108220.
[45] Y. Lin et al, Improving person re-identification by attribute and identity learning, Pattern Recognit. 95 (2019) 151–161.
[46] P. Sudowe, H. Spitzer, B. Leibe, Person Attribute Recognition with a Jointly-Trained Holistic CNN Model, Proc. IEEE Int. Conf. Comput. Vis. vol. 2016-Febru (2016) 329–337.
[47] D. Li, X. Chen, K. Huang, Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios, in: Proc. - 3rd IAPR Asian Conf. Pattern Recognition, ACPR 2015, pp. 111–115, 2016.
[48] A.H. Abdulnabi, G. Wang, J. Lu, K. Jia, Multi-Task CNN Model for Attribute Prediction, IEEE Trans. Multimed. 17 (11) (2015) 1949–1959.
[49] L. Yang, L. Zhu, Y. Wei, S. Liang, P. Tan, Attribute recognition from adaptive parts, in: Br. Mach. Vis. Conf. 2016, BMVC 2016, vol. 2016-September, pp. 81.1-81.11, 2016.
[50] A. Diba, A.M. Pazandeh, H. Pirsiavash, L. Van Gool, DeepCAMP: Deep Convolutional Action & Attribute Mid-Level Patterns, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2016-December (2016) 3557–3565.
[51] J. Zhu, S. Liao, D. Yi, Z. Lei, S. Z. Li, Multi-label CNN based pedestrian attribute learning for soft biometrics, in: Proc. 2015 Int. Conf. Biometrics, ICB 2015, pp. 535–540, 2015.
[52] X. Liu, H. Zhao, HydraPlus-Net: Attentive Deep Features for Pedestrian Analysis, 2017 IEEE International Conference on Computer Vision (ICCV), 2017.
[53] Z. Tan, Y. Yang, J. Wan, H. Hang, G. Guo, S.Z. Li, Attention-based pedestrian attribute analysis, IEEE Trans. Image Process. 28 (12) (2019) 6126–6140.
[54] L.C.B. Haitian Zeng, H. Ai, Z. Zhuang, Multi-task learning via co-attentive sharing for pedestrian attribute recognition, in: 2020 IEEE International Conference on Multimedia and Expo (ICME), 2020, pp. 1–6.
[55] N. Sarafianos, X. Xu, I.A. Kakadiaris, Deep imbalanced attribute classification using visual attention aggregation, ECCV 11215 (2018) 708–725.
[56] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable Person Re-identification: a Benchmark, 2015 IEEE International Conference on Computer Vision (ICCV), 2015.
[57] M. Jaderberg, "Spatial Transformer Networks, in: NIPS'15 Proceedings of the 28th International Conference on Neural Information Processing Systems, 2015, pp. 2017–2025.
[58] Y. Deng, P. Luo, C.C. Loy, X. Tang, Pedestrian attribute recognition at far distance, in: MM '14 Proceedings of the 22nd ACM international conference on Multimedia, 2014, pp. 789–792.
[59] M. Halstead, S. Denman, S. Sridharan, C. Fookes, Locating people in video from semantic descriptions: a new database and approach, Proc. - Int. Conf. Pattern Recognit. (2014) 4501–4506.
[60] R.Y. Tsai, An Efficient and Accurate Camera Calibration Technique for 3D Machine Vision, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1986, pp. 364–374.