Person Retrieval in Surveillance Video using Height, Color and Gender

Anonymous AVSS submission

Paper ID 0003

Abstract

A person is usually described by attributes like height, build, cloth color, cloth type, and gender. Such attributes are known as soft biometrics and they bridge the semantic gap between human description and person retrieval system. This paper proposes a deep learning-based approach for person retrieval using most discriminative soft biometrics namely height, cloth color, and gender. The proposed approach uses Mask R-CNN for pixel-wise person segmentation to remove background clutter and it provides precise boundary around the subject. Color and gender models are finetuned using AlexNet. The algorithm is tested on SoftBioSearch dataset and it achieves good accuracy for person retrieval using the semantic query in challenging conditions.

1. Introduction

Surveillance systems are deployed at many places for security. Currently, person identification is done manually by searching through video footage. Its primary goal is to localize the person of interest using description. For example, a tall male with a black t-shirt and blue jeans. Such descriptions are easy to understand and commonly used by an eye-witness to describe the person. However, descriptors cannot be fed to an automatic person retrieval system and they should be converted to a compatible representation.

The task of person retrieval in the video is very challenging due to occlusion, light condition, camera quality, person's pose, and zoom. However, descriptors like height, cloth color, gender can be deduced from low-quality surveillance video at a distance without cooperation from the subject. Such attributes are known as soft biometrics [1]. A study [2] identified 13 soft biometric traits. A single soft biometric cannot identify the individual uniquely. Therefore, it is important to find multiple and most discriminative soft biometrics for semantic query-based person retrieval.

Person retrieval using semantic description has been widely studied in recent years. Jain et. al. [3] proposes the integration of ethnicity, gender and height to improve the performance of the traditional biometric system. Park et. al. [4] develops visual search engine using dominant color, build and height to find the person of interest. Techniques in [5 – 8] use soft biometrics for person re-identification which aims at spotting a person of interest in other cameras. The appearance-based model proposal by Farenzena et al. [5] uses three complementary aspects of the human appearance; the overall chromatic content, the spatial arrangement of colors into stable regions, and the presence of recurrent textures. Bazzani et al. [6] suggest an appearance-based model that incorporates a global and local statistical feature in the form of an HSV histogram and epitomic analysis respectively. However, techniques [5 – 8] are unsuitable for automatic retrieval as the person is not previously observed.

Description based person retrieval [9 – 11] uses an avatar generated from a semantic description. Avatar incorporates height and clothing color and searches in the image is driven by particle filter [9]. Denman et. al. [10] generates a search query in the form of channel representation using height, dominant color (torso and leg), and clothing type (torso and leg).

Convolutional Neural Network (CNN) based approaches [12 – 14] are becoming popular for person attribute recognition. Multi-label convolutional neural network (MLCNN) [12] predicts gender, age and clothing together with pedestrian attributes. Person's full body image, covered by bounding box, is split in to 15 overlapping 32×32 sized parts are filtered and combined in the cost layer. Dangwei Li et. al. [14] studies the limitation of hand crafted features (e.g., color histograms) and focuses on relationship between different traits. They propose deep learning based single attribute recognition model (DeepSAR) to identify each trait independently and recognize multiple attributes jointly (DeepMAR) to exploit the relationship among the traits.

The approaches based on avatar [9, 10], channel representation [11] and CNN [12 – 14] considers potential area as bounding box around the person. This creates following problems:

1. The bounding box with cluttered background may impact the person detection accuracy specifically for low resolution video, occlusion and multiple views.
2. Person's height is estimated based on bounding box. But, it may be larger in size than a person. This results in to additional rows above the head and below the feet of the person in the image. This leads to incorrect estimation of real world height of the person and also clutter the background.

This paper proposes a deep learning-based approach and uses height, cloth color and gender for retrieval as they are discriminative and commonly occurring descriptors in investigation report. Person's height is view and distance invariant and can be extracted using calibrated camera. The use of color [15] has following advantages: 1. Color sensitivity is independent of direction, view angle and resolution; 2. Color also provides better immunity against noise. Height of the person is estimated using given camera calibration parameters and cloth color and gender are detected using CNN. Main contributions of this work are:

1. Use of semantic segmentation (pixel level segmentation) [16] to detect a person which has following advantages:
   a. It removes unwanted background clutter.
   b. The precision of the segmented boundary yields accurate head and feet points. This results in to better estimate of the real-world height.
   c. It extracts precise patch for torso color classification.
2. The height can also be used to discriminate between the up-right and sitting position of the person. This reduces the search space for the person of interest in up-right position.
3. The paper proposes a generalize framework to deal with low resolution, view and pose variance.

The remainder of the paper is organized as follows. Section 2 describes the proposed approach including details of models used for training data and fusion of various soft biometrics traits to localize the person. Section 3 discusses the test results and accuracy of the approach. Section 4 concludes the paper and discusses possible future work.

2. Proposed approach

This section introduces person retrieval based on height, cloth color and gender. Figure 1 illustrates flow diagram of the proposed framework. The video frame is given to state-of-the-art Mask R-CNN [16] for pixel wise segmentation of persons. Head and foot points are extracted for all segmented persons. Using camera calibration parameters their respective height is computed which is compared with height given in sematic query. Thus, height acts as a filter to reduce the number of persons in the frame. In case of multiple matches, further filtering is done using torso color. The use of semantic segmentation allows background free extraction of color patch from torso. The number of subjects is further reduced by matching of color in semantic query with classified color of extracted patches. The exactness of the final output is improved by using gender classification. Next subsections describe process of height, color and gender estimation for person retrieval.

Height estimation: Person height is view invariant and it also helps to discriminate between the up-right and sitting position of the person. Height is estimated using Tsai camera calibration approach [20], which translates the image coordinate to real world coordinate. SoftBioSearch data set [10] provides 6 calibrated cameras for calculation of real world coordinates. Person's head and feet points are estimated from the semantic segmentation (Figure 1). Steps for height estimation are as follows:

1. Intrinsic parameters matrix ($\boldsymbol{k}$), rotation matrix($\boldsymbol{R}$) and translation vector ($\boldsymbol{t}$) are calculated from given calibration parameters.
2. Perspective transformation matrix is calculated, $\boldsymbol{C} = \boldsymbol{k}[\boldsymbol{R}|\boldsymbol{t}]$.
3. Head and foot points are undistorted using radial distortion parameters.
4. World coordinate for feet is set as $Z = 0$ *and* X, Y world coordinates are derived using inverse transformation of $\boldsymbol{C}$.
5. Use X, Y coordinate (of step-4) to calculate Z coordinate of the head which also represent height.

Estimated height helps in reducing the search space within test frame based on description (e.g., 150 – 170 cm). Test frame would now contain only the person(s) with matching height description.

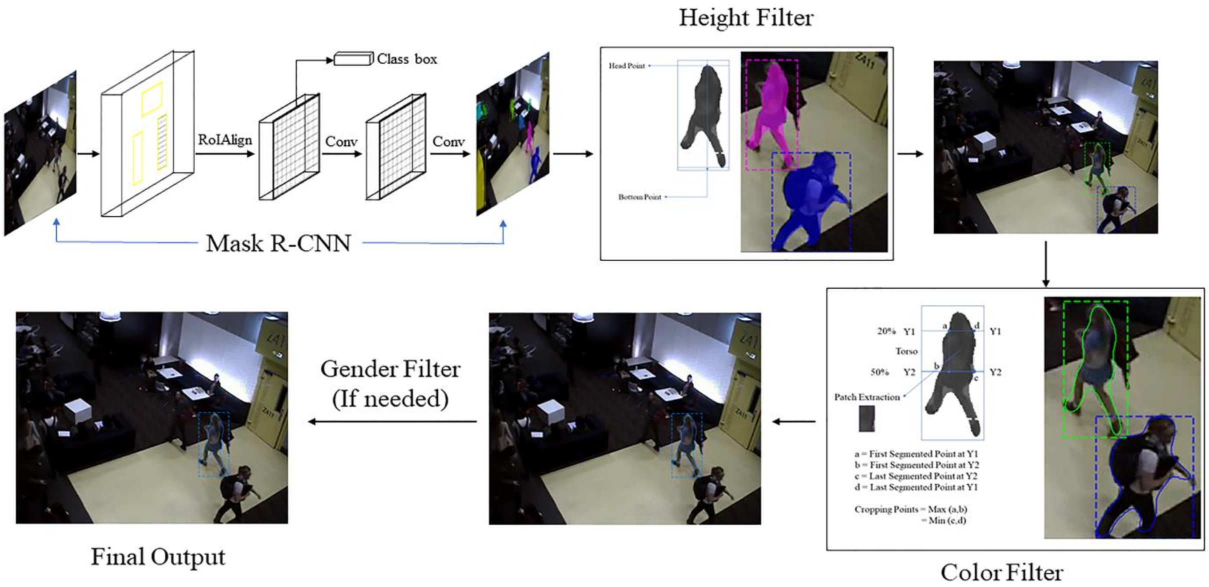


Figure 1: Proposed approach of person retrieval using height, cloth color and gender.

During training, annotated head and foot points are used to compute height from all frames of the video sequence. The height is then averaged (H_{avg}) over all frames. Over the same training sequence, it was observed that average height computed from automated head and foot point is larger than H_{avg}. This difference yields wrong estimation therefore; it is subtracted from average estimated height during testing to normalize the error.

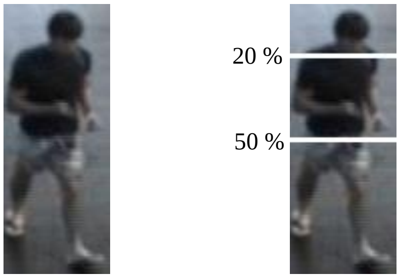


Figure 2: Extraction of torso and leg region from body

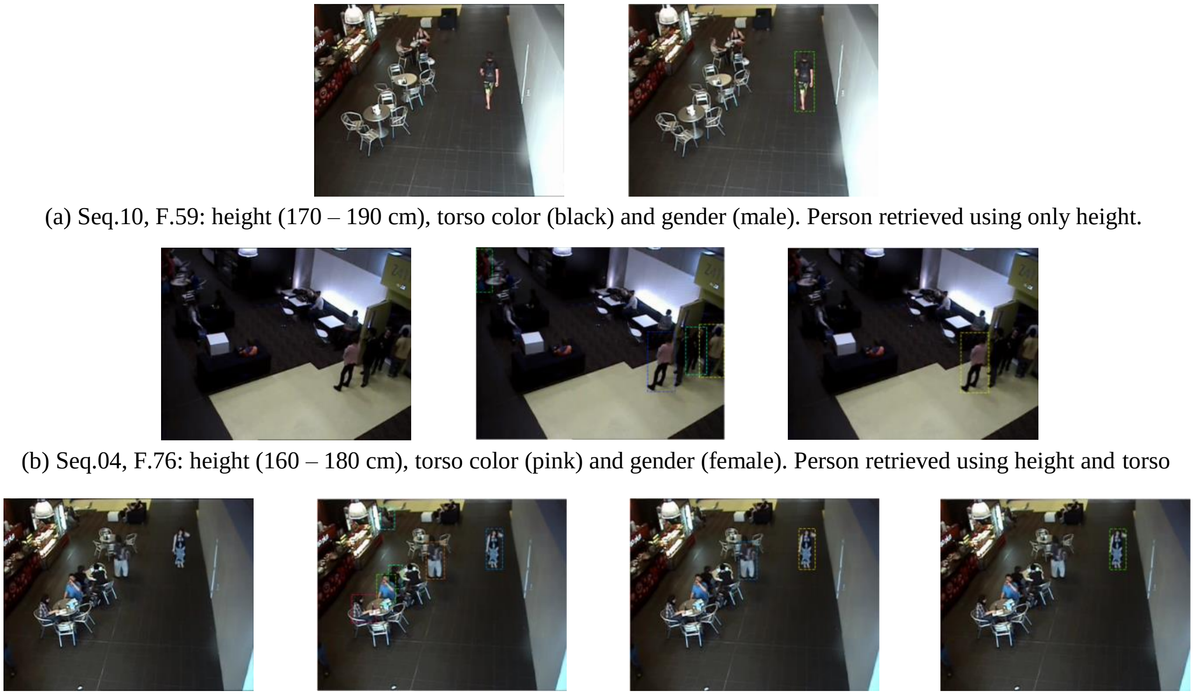
Torso color prediction: Mask R-CNN generates class label with probability score and semantic segmentation. The torso and leg regions are extracted using golden ratio for height. The upper 20-50% portion is classified as torso, while the lower 50-100% represents legs of the person. The torso and leg segmentation is shown in Figure 2. The use of semantic segmentation allows extraction of color patch from precise torso region. This is shown in Figure 1 (ref. color filter) as the region bounded by points ‘a’, ‘b’, ‘c’ and ‘d’. This removes unwanted background clutter in torso region improving color classification. The extracted color patch is then passing to fine-tuned AlexNet [18], which predicts probability score. The dataset annotations contain two colors *Torso Color* and *Torso Second Color* for each subject. In case of multiple matches due to *Torso Color*, the algorithm will refine the result using *Torso Second Color* if present. This feature helps to narrow down the search space.

Gender classification: The proposed algorithm accurately retrieves the person for most cases using height and cloth color. But, in case of multiple matches the algorithm uses gender classification. The AlexNet is fine-tuned using full body images for male and female gender classification.

2.1. Implementation details

This section covers details about dataset, data augmentation and AlexNet training. The proposed approach uses the pretrained weights of Microsoft COCO [17] for detection and semantic segmentation. It has the Average Precision (AP) of 35.7 on COCO test set.

Dataset overview: This paper uses the SoftBioSearch database [10] which includes 110 unconstrained training video sequences, recorded using 6 stationary calibrated cameras. Each of the sequences contains 16 annotated soft biometric traits and 9 body markers for subject localization. The 9 body markers are top of the head, left and right neck, left and right shoulder, left and right waist, approximate toe position of the feet. The test dataset contains video sequences of 41 subjects with semantic query. The sequence length has 21 to 290 frames for training subjects. Discarding the frames with partial occlusion, the resulting training set has

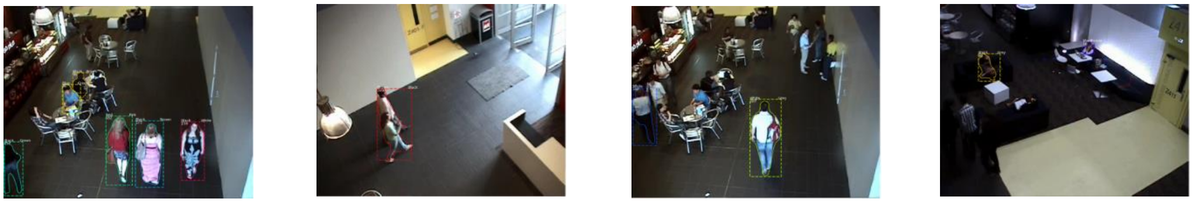


(a) Seq.10, F.59: height (170 – 190 cm), torso color (black) and gender (male). Person retrieved using only height.

(b) Seq.04, F.76: height (160 – 180 cm), torso color (pink) and gender (female). Person retrieved using height and torso

(c) Seq.8, F.31: height (130 – 160 cm), torso color (pink) and gender (female). Person retrieved using height, torso color and gender.

Figure 4: True positive cases of person retrieval with semantic query.



(a) (b) (c) (d)

Figure 5: True negative cases of person retrieval. (a) incorrect color classification with multiple persons, (b) multiple persons with occlusion, (c) multiple persons with same torso color and height class and (d) person detection fails.

8577 images from 110 subjects.

Data augmentation: It is a practice in deep learning to augment the training samples for improvement in performance and robustness. E.g., training AlexNet with 8577 images may result in to overfitting, which is avoided using data augmentation. Each train image is horizontally and vertically flipped, rotated with 10 angles {1^o, 2^o, 3^o, 4^o, 5^o, -1^o, -2^o, -3^o, -4^o, -5^o} and brightness increased with gamma factor of 1.5.

2.2. AlexNet training

The training is accomplished on workstation with Intel Xeon core processor and accelerated by NVIDIA Quadro K5200 of 8 GB GPU. Color and gender models are finetuned using AlexNet which is pretrained on ImageNet [19] dataset.

Color training: The SoftBioSearch [10] database contains 1704 color patches divided in to 12 culture colors. Additional patches for color training are extracted using 4 body markers namely left and right shoulder, left and right waist from the training dataset. In order to deal with illumination changes these patches are augmented by increasing brightness with gamma factor of 1.5. This generated approximately 17000 color patches. The last four layers of AlexNet namely (Conv5, fc6, fc7, and fc8) are fine-tuned for color training. It is trained for 30 epochs with learning rate of 0.001, dropout set to 0.50 and effective batch size of 128.

Gender training: The data augmentation generated 105980 images for gender training which is approximately 13 times larger than original training set (8577). Gender training is accomplished by fine tuning last 3 layers (fc6, fc7, and fc8) of AlexNet. The model is fine-tuned for 20 epochs. The learning rate is set to 0.0001 and training batch size has 64 images. A minor over-fitting is controlled by increasing the dropout rate to 0.40. The overall fine-tuning process took 16 – 18 hours for completion.

3. Experimental results:

This section covers the qualitative experimental results derived on test data set of 41 subjects. The ground truth (i.e., person of interest) is established by manually mapping semantic test query to video frames of the test set. The ground truth is established after first 30 initialization frames of each subject. Figure 4 shows true positive cases of the person retrieval using semantic queries. For example, Seq.10, F.59 indicates sequence number 10 with frame number 59 in the test set. Frames from left to right indicate input test frame, output of height filter, output of color filter and gender filter respectively. Figure 4(a) shows person of Seq.10, F.59 with semantic query height (170 – 190 cm), torso color (black) and gender (male). Person is retrieved using only single biometric trait i.e., height. Person of Seq.4, F.76 with semantic query height (160 – 180 cm), torso color (pink) and gender (female) is shown in Figure 4(b). It can be observed that multiple persons are detected with same height class (ref. middle image in 4(b)). The correct person of interest is retrieved by adding torso color query to the height. Figure 4(c) shows person of Seq.8, F.31 with semantic query height (130 – 160 cm), torso color (pink) and gender (female) in which algorithm utilizes all 3 semantic queries i.e., height, torso color and gender to retrieve person of interest. Thus, algorithm retrieve person with rank-1 match by utilizing minimum number of soft biometric traits and in case of multiple match use additional soft biometrics to retrieve unique match.

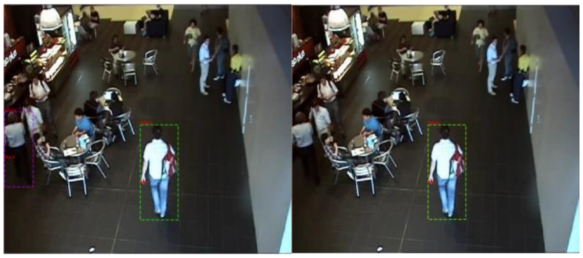


Figure 6: Use of leg color for person retrieval

Figure 5 shows the results where the algorithm fails to retrieve the person correctly. It also indicates challenging conditions in the test dataset. Figure 5(a) shows person of Seq.16, F.31 with semantic query height (160 – 180 cm), torso color (pink) and gender (female). It shows incorrect torso color classification with pink color classified as black (*Torso Color*) and brown (*Torso Second color*). It could be due to presence of brown hair at the back side of person. Figure 5(b) contains occlusion due multiple persons and as a result Mask R-CNN creates single bounding box for Seq.25, F.34 with semantic query height (150 – 170 cm), torso color (green) and gender (female). Algorithm fails to retrieve person uniquely when multiple persons with same torso color and same height class are present, e.g., Seq.18, F.31 with semantic query height (180 – 210 cm), torso color (white) and gender (male) (Ref. Figure 5(c)). Figure 5(d) (Seq.1, F.73) shows the scenario where the person of interest is merged with black background (except black and white torso region) due to poor illumination and person detection fails in this frame.

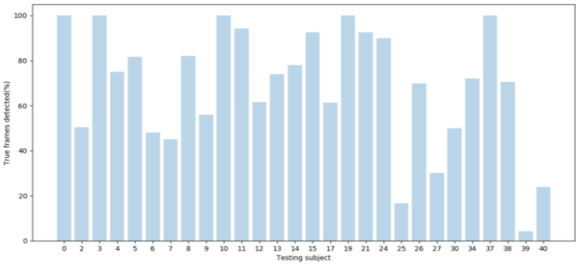


Figure 7: True Positive (TP) rate of person retrieval.

True Positive (TP) rate is calculated for testing dataset of 41 persons as a qualitative measure of accuracy for the proposed approach. It is computed as follows.

$$TP\ (\%) = \frac{No.of\ frames\ in\ which\ person\ retrieved\ correctly}{Total\ No.of\ frames}$$

The algorithm is able to retrieve 28 persons correctly with varying TP rate. Figure 7 shows the % TP rate (Y axis) and index of correctly retrieved person (X axis). Average TP rate of the algorithm for 28 correctly retrieved person is 65.8%. Among 28 persons, 19 are retrieved with TP rate more than 60%, 5 with TP rate between 30% - 60% and 4 persons with TP rate between 0% - 30%. For 19 persons (TP rate > 60%), height and color model works extremely well. The TP rate is very poor for some sequences. For example, in Seq.39 torso color (brown) is incorrectly classified. This is due to low number of color patches used in training and therefore, model is unable to classify true color. The color classification can be improved by adding more color patches. In many frames, person of interest is occluded e.g., Seq. 25 and semantic segmentation could not extract precise boundary around the person. Similarly, in Seq.1 person of interest is merged with background due to poor illumination resulting in poor detection. Algorithm achieves average TP rate of 52.2% considering all 41 test persons.

4. Conclusion

The proposed approach retrieves the person in surveillance video using a semantic query based on height, cloth color and gender. Use of semantic segmentation allows better height estimation and precise color patch extraction from the torso. The algorithm correctly recovers 28 persons out of 41 in very challenging dataset with soft biometric traits. Average accuracy of the algorithm is 52.2%. Problem described in Figure 5(c), where algorithm fails to retrieve person, can be resolved by incorporating additional soft biometric in retrieval process. E.g., Figure 6 shows the improved result where the person is correctly retrieved using leg color. Future work will focus on improving results by incorporating other soft biometric traits like torso texture, body accessory and investigate the mechanism to make the proposed approach more robust for person retrieval.

References

[1] A. Dantcheva, P. Elia, A. Ross. What Else Does Your Biometric Data Reveal? A Survey on Soft Biometrics. IEEE Transactions on Information Forensics and Security 11(3): 441-467, 2015.

[2] M. D. MacLeod, J. N. Frowley and J. W. Shepherd. Whole body information: Its relevance to eyewitnesses. In D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), Adult eyewitness testimony: Current trends and developments, pp. 125 – 143. New York, US: Cambridge University Press 1994.

[3] A. K. Jain, S. C. Dass, and K. Nandakumar. Soft biometric traits for personal recognition systems. In International Conference on Biometric Authentication, Hong Kong, 2008, pp. 731–738.

[4] U. Park, A. Jain, I. Kitahara, K. Kogure, and N. Hagita. Vise: Visual search engine using multiple networked cameras. In 18th International Conference on Pattern Recognition (ICPR), 2006, pp. 1204 –1207.

[5] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 2360–2367.

[6] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino. Multiple-shot person re-identification by HPE signature. In 20th International Conference on Pattern Recognition (ICPR), 2010, pp. 1413–1416.

[7] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2528 – 2535.

[8] Y. Xu, L. Lin, W.-S. Zheng, and X. Liu. Human re-identification by matching compositional template with cluster sampling. In IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3152 – 3159.

[9] S. Denman, M. Halstead, A. Bialkowski, C. Fookes and S. Sridharan. Can you describe him for me? A technique for semantic person search in video. In International Conference on Digital Image Computing: Techniques and Applications, (DICTA), 2012, pp.1–8.

[10] M. Halstead, S. Denman, C. Fookes and S. Sridharan, Locating people in video from semantic descriptions: a new database and approach. In International Conference on Pattern Recognition (ICPR), 2014 pp. 4501 – 4506.

[11] S. Denman, M. Halstead, C. Fookes and S. Sridharan. Searching for people using semantic soft biometric descriptions. Pattern Recognition Letters, 68 (Part 2), pp. 306-315, 2015.

[12] J. Zhu, S. Liao, D. Yi, Z. Lei and S. Z. Li. Multi-label CNN Based Pedestrian Attribute Learning for Soft Biometrics. In International Conference on Biometrics (ICB), 2015, pp. 535 – 540.

[13] P. Sudowe; H. Spitzer and B. Leibe. Person Attribute Recognition with a Jointly-Trained Holistic CNN Model. In IEEE International Conference on Computer Vision Workshop (ICCVW), 2015, pp. 329 – 337.

[14] D. Li, X. Chen and K. Huang. Multi-attribute Learning for Pedestrian Attribute Recognition in Surveillance Scenarios. In IAPR Asian Conference on Pattern Recognition (ACPR), 2015, pp. 111 – 115.

[15] P. Shah, M. S. Raval, S. Pandya, S. Chaudhary, A. Laddha and H. Galiyawala. Description Based Person Identification: Use of Clothes Color and Type. In National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), Dec – 2017, IIT Mandi.

[16] He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. In IEEE International Conference on Computer Vision (ICCV), 2017,Oct 22 (pp. 2980-2988).

[17] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft coco: Common objects in context. In European conference on computer vision 2014 Sep 6 (pp. 740-755). Springer, Cham.

[18] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems 2012 (pp. 1097-1105).

[19] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on 2009 Jun 20 (pp. 248-255).

[20] R. Tsai, “A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses”, IEEE Journal of Robotics and Automation, vol. 3, pp 323-344, 1987.