Region-Based CNNs for Pedestrian Gender Recognition in Visual Surveillance Environments

Ehsan Yaghoubi*, Pendar Alirezazadeh*, Eduardo Assunção*, João C. Neves† and Hugo Proença*
*IT: Instituto de Telecomunicações, University of Beira Interior, Faculty of Computer Engineering, Covilha, Portugal
†TOMI: City's Best Friend, Viseu, Portugal

***Abstract*—Inferring soft biometric labels in totally uncontrolled outdoor environments, such as surveillance scenarios, remains a challenge due to the low resolution of data and its covariates that might seriously compromise performance (e.g., occlusions and subjects pose). In this kind of data, even state-of-the-art deep-learning frameworks (such as ResNet) working in a holistic way, attain relatively poor performance, which was the main motivation for the work described in this paper. In particular, having noticed the main effect of the subjects' "*pose*" factor, in this paper we describe a method that uses the body keypoints to estimate the subjects pose and define a set of regions of interest (e.g., *head*, *torso*, and *legs*). This information is used to learn appropriate classification models, specialized in different poses/body parts, which contributes to solid improvements in performance. This conclusion is supported by the experiments we conducted in multiple *real-world* outdoor scenarios, using the data acquired from advertising panels placed in crowded urban environments.**

***Index Terms*—Pedestrian attribute recognition, skeleton detection, pose estimation, segmentation.**

I. Introduction

Being often the first mentioned attribute to describe a person, gender estimation is useful in many areas of computer vision, such as surveillance, forensic affairs, marketing, and human-robot interaction. In the first decade of this century, datasets were small and most approaches were based on handcrafted features such as Histogram of Oriented Gradients (HOG). However, after the advent of deep learning frameworks, scholars focused on collecting extensive labeled data and developing deeper networks.

In the literature, gender estimation from facial images has received more attention than whole-body. However, in this paper, we use full-body images since in Pedestrian Attribute Recognition (PAR) scenarios not only the quality of facial regions decreases, but also the body features are more robust to far distances.

[1] proposes a fine-tuned CNN model to predict the gender from the "front", "back" and "both" views. They employ a parsing mechanism via the decompositional neural network (DNN) to remove the background. The foreground is then parsed in the upper and lower bodies so that the two CNNs are fine-tuned. As a conclusion, feeding upper-body images to the network slightly improves the results. However, they have gray scaled and forced-squared the images which cause the loss of color-based features and data deformation. In [2], authors apply HOG alongside a CNN and concatenate the extracted features that are used as the input of a Softmax classifier. Although the expressiveness of the data is protected in this method, feature redundancy in the last layer can lead to a biased model that degrade the performance in real-world applications. [3] presents another work that adopts an extra thermal camera for data acquisition. Using CNN methods, they extract the features from visible images and thermal maps and fuse them in score level by exploiting Support Vector Machine (SVM) learner. As they apply thermal images for recognition, the algorithm can fail in crowded places with occlusion, which is a real and critical scenario.

In addition to the mentioned weaknesses, the datasets in previous works are mainly collected from one location which can cause some easiness such as: monotonous illumination, stable camera settings, controlled occlusion, similar background, and controlled distance acquisition. While in this paper, we collect a dataset from outdoor and indoor advertisement panels in more than 100 cities of Portugal and Brazil*.

Further, we propose a Pedestrian Gender Recognition Network (PGRN) which provides several decisions based on the subject pose and some Regions of Interest (RoI) so that the decision with maximum certainty is reported as the final recognition (Fig. 1). The performed experiments on three datasets show the superiority of the proposed algorithm in comparison with the state-of-the-art methods, as detailed in section 3.

II. Pedestrian Gender Recognition Network (PGRN)

Regarding the impact of pose variation on the biometric system performance, we develop our proposed algorithm on a human body keypoint detection and tracker platform. In general, the suggested PGRN is divided into the following steps: training the baseline network called Base-Net, key point detection and tracking, pose extraction, RoI extraction, fine-tuning Pose-Sensitive Networks (PSN), and score fusion.

This research is funded by the "FEDER, Fundo de Coesao e Fundo Social Europeu" under the "PT2020 - Portugal 2020" program, "IT: Instituto de Telecomunicações" and "TOMI: City's Best Friend". Also, the work is funded by FCT/MEC through national funds and when applicable co-funded by FEDER PT2020 partnership agreement under the project UID/EEA/50008/2019.

*https://tomiworld.com/locations/

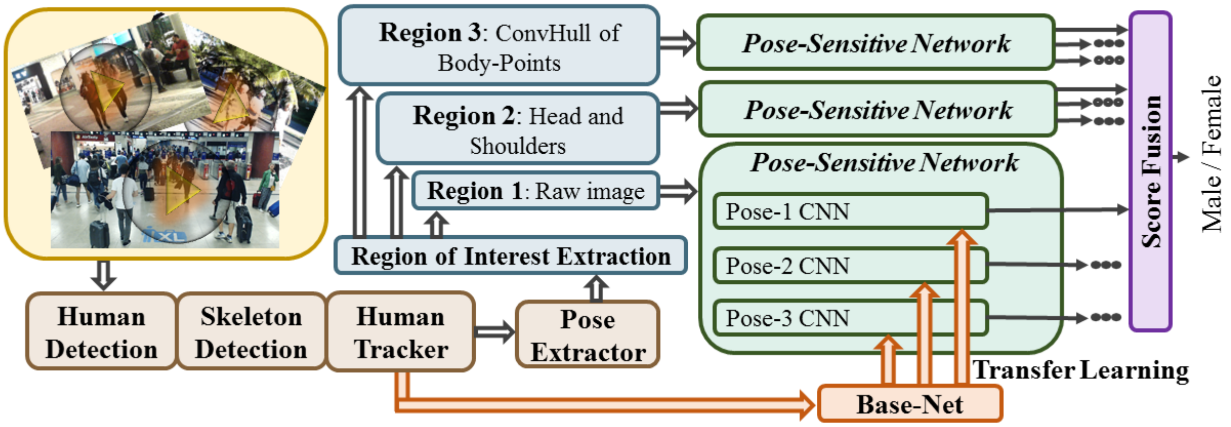


Fig. 1: Overview of the proposed algorithm called PGRN. Taking advantage of the human detector, skeleton detector, and human tracker, we extract the bounding boxes alongside 16 body keypoints for each person. Afterward, the training set is split into three subsets corresponding to the desired poses (i.e. frontal, rear, and lateral). The RoIs are then extracted and fed to the Pose-Sensitive Network (PSN) which is constructed from three specialized ResNet50 networks. The weights of a pre-trained network (i.e. Base-Net) are shared with each of these PSNs to reduce the time of training. Finally, the most confident score from the RoIs is considered as the final score for recognition.

A. *Base-Net*

Although the pre-trained CNNs on the ImageNet dataset have shown promising results on various recognition tasks, it is interesting to note that training from scratch or updating the weights of all layers necessarily leads to better results upon the availability of sufficient data. As we have collected a large proprietary dataset (i.e. Biometria e Deteção de Incidentes (BIODI)), the weights of the network trained on the ImageNet dataset are considered as the initial weights for our model. Afterward, the whole layers of the network are trained on raw images of the BIODI. This network is named as Base-Net that later will be used for transferring the knowledge to the PSNs.

B. *Body Key-Point Detection and Tracking*

BIODI is composed of 216 video clips of wild visual surveillance environments taken from different countries. We started by analyzing each video using a state-of-the-art approach called Alphapose [4] that is an accurate real-time and multi-person skeleton detector based on an object detection method named Faster-RCNN [5]. This object detector provides the bounding boxes (BBs) of multiple humans in each frame. Then, the human BBs are fed to the Spatial Transformer Network (STN) [6], which yields high quality dominant human proposals. In other words, the out put of the STN are some transformed human proposals, therefore, after estimating the skeleton of each person using the Single Person Pose Estimator (SPPE) [7], each set of the body keypoints needs to be mapped to the original image coordinate using a de-transformer network.

So far, the detection of BBs and skeleton of each person in each frame is done. To perform the tracking, the straight forward approach is to connect the current skeletons to the closest skeletons in the next frame. However, this method produces errors when there are several poses close to each other. Therefore, we apply Poseflow [8] that works based on a small inter-frame skeleton distance (d_c) and a large intra-frame skeleton distance (d_f) of the form Eq. 1. Finally, we storage all the BBs and body keypoints related to each human subject to the disk for the next step.

$$
\begin{aligned}
&d_c(S^{(1)},\, S^{(2)}) = \sum_{n=1}^{N} \frac{f_2^n}{f_1^n}, \\
&d_f(S_1,\, S_2|\,\{\sigma_1, \sigma_2, \lambda\}) = \frac{1}{K_{sim}(S_1, S_2|\sigma_1)} + \frac{\lambda}{H_{sim}(S_1, S_2|\sigma_2)}, \\
&s.t.\;\; K_{sim}(S_1,\, S_2|\sigma_1) = \\
&\qquad\qquad \begin{cases} \sum_{n=1}^{N} \tanh\frac{c_1^n}{\sigma_1} . \tanh\frac{c_2^n}{\sigma_1} & ;\text{ if }\; S_2^n \;\text{ in }\; B(S_1^n) \\ 0\,;\ \text{Otherwise}, \end{cases} \\
&s.t.\;\; H_{sim}(S_1,\, S_2|\sigma_2) = \sum_{n=1}^{N} e^{-\frac{(S_1^n - S_2^n)}{\sigma_2}},
\end{aligned}
\tag{1}
$$

where S_1 and S_2 are two skeletons related to two different individuals in a frame in $B(S_1^n)$ and $B(S_2^n)$ bounding boxes, respectively. f_1^n and f_2^n are extracted features of these boxes and $n \in \{1, ..., N\}$ in which N represents the number of body keypoints, and σ_1, σ_2, and λ can be determined in a data-driven manner.

C. *Pose Inference*

For a biometric system specialized in specific human body-pose, various body gestures provide different features, therefore, unseen poses in the test set highly impact its performance. On the other hand, pose-specialized networks are not able to learn the important features if we split the train set to many subsets of different poses. Regarding this matter and number of images of our dataset, we considered only the three most common poses of pedestrians, including ”frontal”, ”rear”, and ”lateral” views.

As the BBs are extracted using an object detector, the aspect ratio (width/height) of each BB is 1.75. We visualized quite a few numbers of body keypoints (see Fig. 2(a)) on the resized images (175x100) and discovered that individuals with shoulder-width lower than nine pixels (out of 100 pixels) in the invariant-scale RoIs can be a nominate for lateral view images. It worth mentioning that, we considered the other body keypoints to perform this experiment, however, the best results are obtained using the shoulder-width points. If $p_i = (x_i, y_i)$ represents the coordinates of body points, the desired poses are:

$$\text{Pose} \equiv \begin{cases} \text{Frontal view;} & \text{if } x_v - x_z < -9 \\ \text{Rear view;} & \text{if } x_v - x_z > 9 \\ \text{Lateral view;} & \text{if } |x_v - x_z| =< 9 \text{ pixels,} \end{cases} \quad (2)$$

where (x_v, y_v) and (x_z, y_z) respectively are 13^{th} and 14^{th} body-point coordinates illustrated in Fig. 2(a).

D. RoI: Segmentation and Cropping Strategies

By joining the exterior body points p we obtain a polygon, we find it useful to create a mask by applying Convex-Hull on this set. For N points $p_1, ..., p_N$, the Convex-Hull is the set of all convex combinations of its points such that in a convex combination each point has a positive weight w_i. These weights are used to compute a weighted average of the points. For each choice of weights, the obtained convex combination is a point in the Convex-Hull. Therefore, choosing weights in all possible ways, we can form a black polygon-shape as Fig. 2(b). In a single equation, the Convex-Hull is the set:

$$CH(N) = \left\{ \sum_{i=1}^{N} w_i p_i : w_i \geq 0 \, \text{for all } i, \text{ and } \sum_{i=1}^{N} w_i = 1 \right\} (3)$$

Figure 2 illustrates this process for a sample image. To avoid information lost when performing the Convex-Hull algorithm, we consider two extra points (x_l, y_l) and (x_r, y_r) near the ears. Therefore, $y_l = y_r = \frac{y_n + y_h}{2}$ and $x_l = x_n - y_l$, $x_r = x_n - y_r$, where (x_n, y_n) and (x_h, y_h) are 9^{th} and 10^{th} body-point coordinates illustrated in Fig. 2 (a), respectively. The polygon-mask is then produced by painting inside of the obtained Convex-Hull with black, and this mask is employed to segment the raw images.

Considering that the facial region carries information about most human traits, including gender, we used different sets of body points such as the elbow, chest-bone, head, neck, and shoulders to crop the head. Under visual inspection, the best results are obtained using the head, chest-bone, and shoulders' points that have been shifted out ten pixels.

E. Pose Sensitive Network and Score Fusion

The Pose Sensitive Network (PSN) is composed of three sub-networks, specialized in three poses (i.e. frontal, rear, and lateral poses). Using weight-sharing, the knowledge of the Base-Net is transferred to these sub-nets. For each image, there are three patches (i.e. head, polygon, whole image) corresponding to three PSNs (see Fig. 1). The obtained scores for each patch are then concatenated, and the highest one is selected as the final score of the image, which means that the model decides based on a optimistic perspective. For example, in case of partial body-occlusion and low score recognition for the full-body image, the model presumably decides based on the head-crop region.

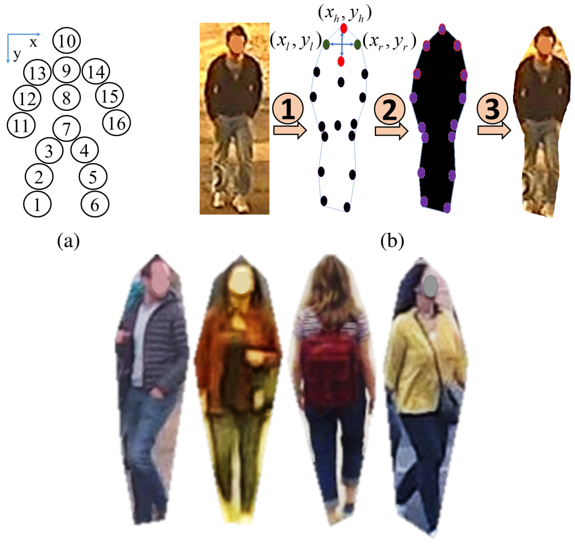


Fig. 2: Foreground segmentation process. After determining the exterior border using the Convex-Hull, a mask is created and the foreground is cropped. (a) Body keypoints (b) Red points are considered as a reference for adding two green points near the head so that the polygon-crop contains the head and hair (c) Samples of segmented images which will have a black background in training phase.

III. Experiments and Discussion

First, we describe the strategy of the data collection and discuss the unique features of the collected dataset. We then briefly explain the two public datasets for which we evaluated our model. Finally, after describing the experimental settings, we provide the results.

A. Datasets

In general, deep-learning-based biometric systems are sensitive to data variability. Due to the environment and subject dynamics, a biometric system trained in a specific place cannot produce the best results in unseen places. This even becomes more critical in universal systems dealing with humans as the subject of interest, because not only the environment alters, but the styles of clothing and body pose differ in various situations. For instance, the recognition rate will be highly affected in a cold and rainy night as people usually cover their bodies, heads, and faces while carrying an umbrella which has occluded the upper body. Therefore, regarding the lack of datasets that cover a wide range of variations in the environment and pedestrian, we collected the BIODI dataset from 36 advertisement panels in Portugal and Brazil at indoor and outdoor locations; different moments of the day including morning, noon, evening and night; and various weathers. Table I summarizes the statistics of this dataset. Each panel has

Factors	Statistics
No. of videos, subjects, and BBs	216, 13876, 503433
Length of videos	7 minutes
Frame rate extraction	7 frames/sec.
Aspect ratio of BBs (Height/Width)	1.75
No. of frontal, rear, and lateral BBs	256485, 235564, 11384

TABLE I: Statistics of the BIODI dataset

one embedded camera with 1.5-meter vertical distance from the ground. Table II shows several samples of the BIODI dataset. It worth mentioning that this private dataset has been annotated manually for 16 soft biometric labels including gender, age, weight, race, height, hair color, hair style, beard, mustache, glasses, head attachments, upper-body cloths, lower-body clothes, shoes, accessories, and action.

To make our results reproducible, we report the performance of our method on public datasets such as PETA (excluding MIT) and MIT. Briefly, MIT pedestrian dataset consists of 888 outdoor images with 64x128 pixels annotated for frontal and rear views. Approximately, half of the images are in frontal view, and female's share is one-third of the dataset. PETA is a collection of 19000 images consisting of 10 different datasets, including the MIT dataset. However, MIT is excluded from PETA since the proposed model will be evaluated on it, separately. It is worth mentioning that, in PETA benchmark, the number of males and females are almost the same and there is no view-wise annotation.

B. Experimental Settings

In our experiments, we use Python 3.5 and Keras 2.1.2 API on top of the Tensorflow 1.13. In order to avoid over fitting, we add the batch normalization, max pooling, and drop out layers to the ResNet50. The learning rate is set to 0.005 for the Stochastic Gradient Descent (SGD) optimizer. It is worth mentioning that we resized the images to 175 x 100 pixels, applied standardization per image, and performed horizontal mirror augmentation.

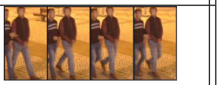
Description	Samples
Outdoor, Noon, Occlusion	
Outdoor, Summer, Night	
Outdoor, Winter, Night	
Outdoor, Fall, Evening	
Outdoor, Summer, morning	
Indoor, Spring, Occlusion	

TABLE II: Sample images of the BIODI dataset that guarantees a wide spectrum of subject and environment changes.

We evaluate the proposed model on three datasets BIODI, MIT, and PETA such that 70% of the BIODI (i.e. 352400 images) and PETA (i.e. 12680 images) datasets are allocated to the training phase. As MIT is a small dataset with 888 images, we used 50% of the data for test phase to have stable results beacuse in each test-run the recognition rate have some variations.

C. Results and Discussion

Considering the explanations in the previous section, the ex periments were conducted in three forms: raw images, head-cropped regions, and polygon-shape regions. Afterward, each trained model is tested. Table III shows the results of the proposed model on the RoIs which indicates that lateral-view state is the most difficult recognizable pose with around 84% and 80% accuracy for the BIODI and PETA datasets, respectively. Furthermore, Frontal-Net outperformed the Base-Net by 1.6% while Rear-Net improved the results from 84.49% to 85.18%, and Lateral-Net estimated the gender slightly better. Moreover, the increase of the 2% accuracy in polygon-crop images shows that the background negatively affects the performance of the networks. Hence, developing the powerful segmentation algorithms for human full-body is suitable for further studies.

Table IV shows the evaluation of the proposed approach on MIT dataset. Notably, we achieve an average accuracy of 90.0%, 87.9%, and 89.0% for the frontal, rear, and mixed-view images, respectively, that are outperforming the results obtained by other methods.

IV. Conclusions and Future Works

Regarding the ubiquitous surveillance cameras and the low-quality facial acquisitions, it is necessary to develop methods that deal with full-body images, occlusions, pose variation, and various illuminates. To this end, we proposed an algorithm for pedestrian gender recognition in crowded urban environments so that the output of a body-joints detector is applied for splitting the images into three common poses. Further, taking advantage of transfer learning, the specialized networks were fine-tuned for extracted RoIs. Extensive experiments on multiple challenging datasets showed that proposed PGRN can effectively estimate the gender and consistently outperform the state-of-the-art methods. As the next step, we have focused on developing an end-to-end network capable of estimating body related soft biometric traits such as weight, age, height, and race.

References

[1] M. Raza, M. Sharif, M. Yasmin, M. A. Khan, T. Saba, and S. L. Fernandes, "Appearance based pedestrians' gender recognition by employing stacked auto encoders in deep learning," *FUTURE GENER COMP SY*, vol. 88, pp. 28–39, 2018.

[2] L. Cai, J. Zhu, H. Zeng, J. Chen, C. Cai, and K.-K. Ma, "HOG-assisted deep feature learning for pedestrian gender recognition," *Journal of the Franklin Institute*, vol. 355, no. 4, pp. 1991–2008, mar 2018. [Online]. Available: https://doi.org/10.1016/j.jfranklin.2017.09.003

Images	Network	**BIODI**	Frontal	Rear	Lateral	**PETA**	Frontal	Rear	Lateral
Raw	Base-Net	85.68	85.96	84.49	79.70	86.77	89.18	89.94	75.99
	Frontal-Net	-	87.53	-	-	-	90.56	-	-
	Rear-Net	-	-	85.18	-	-	-	93.06	-
	Lateral-Net	-	-	-	79.87	-	-	-	77.20
Head	Frontal-Net	-	88.42	-	-	-	88.73	-	-
	Rear-Net	-	-	85.13	-	-	-	90.15	-
	Lateral-Net	-	-	-	78.09	-	-	-	77.37
Polygon	Frontal-Net	-	90.44	-	-	-	91.29	-	-
	Rear-Net	-	-	87.44	-	-	-	91.44	-
	Lateral-Net	-	-	-	80.99	-	-	-	76.06
Fusion	Frontal-Net	-	**92.19**	-	-	-	**92.15**	-	-
	Rear-Net	-	-	**88.86**	-	-	-	**93.58**	-
	Lateral-Net	-	-	-	**84.16**	-	-	-	**80.16**

TABLE III: Accuracy for the experiments on BIODI and PETA in percentage. The experiments on raw, head-crop, and polygon-crop images suggest that head-crop images provide the weakest results and confirm the fact that in surveillance scenarios, full-body recognition is more robust. Secondly, we perceived that as BIODI contains various environments, polygon segmentation provides better results while this is not true for PETA dataset. Finally, the last row of the table indicates that the adopted strategy for score fusion produces the highest score and accuracy among other approaches.

View	[9]	[10]	[11]	[12]	[1]	Proposed Method
Front	76.0	79.5	81.0	82.1	82.9	**90.0**
Back	74.6	84.0	82.7	81.3	81.8	**87.9**
Mixed	-	78.2	80.1	82.0	82.4	**89.0**

TABLE IV: Results on MIT test set in percentage.

[3] D. Nguyen, K. Kim, H. Hong, J. Koo, M. Kim, and K. Park, “Gender recognition from human-body images using visible-light and thermal camera videos based on a cnn for image feature extraction,” *Sensors*, vol. 17, no. 3, p. 637, mar 2017. [Online]. Available: https://doi.org/10.3390/s17030637

[4] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, “RMPE: Regional multi-person pose estimation,” in *ICCV*, 2017.

[5] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *NIPS*, 2015, pp. 91–99.

[6] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, “Spatial transformer networks,” in *Advances in neural information processing systems*, 2015, pp. 2017–2025.

[7] A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” in *ECCV*. Springer, 2016, pp. 483–499.

[8] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu, “Pose Flow: Efficient online pose tracking,” in *BMVC*, 2018.

[9] L. Cao, M. Dikmen, Y. Fu, and T. S. Huang, “Gender recognition from body,” in *Proceedings of the 16th ACM international conference on Multimedia*. ACM, 2008, pp. 725–728.

[10] G. Guo, G. Mu, and Y. Fu, “Gender from body: A biologically-inspired approach with manifold learning,” in *ACCV*. Springer, 2009, pp. 236–245.

[11] C. D. Geelen, R. G. Wijnhoven, G. Dubbelman *et al.*, “Gender classification in low-resolution surveillance video: in-depth comparison of random forests and svms,” in *VSTIA2015*, vol. 9407. International Society for Optics and Photonics, 2015, p. 94070M.

[12] M. Raza, C. Zonghai, S. Rehman, G. Zhenhua, W. Jikai, and B. Peng, “Part-wise pedestrian gender recognition via deep convolutional neural networks,” in *2nd IET ICBISP*. Institution of Engineering and Technology, 2017. [Online]. Available: https://doi.org/10.1049/cp.2017.0102