Semantic segmentation of clothes in the context of soft biometrics using deep learning methods

Andrei de Souza Inácio, Anderson Brilhador, Heitor Silvério Lopes
Graduate Program in Electrical Engineering and Industrial Informatics (CPGEI)
Federal University of Technology - Paraná
Curitiba, Brazil
{andrei.inacioo, andersonbrilhador, heitorslopes}@gmail.com

***Resumo*—Soft biometrics is an emerging area of research, mainly due to its large applicability in people surveillance. It is related to human characteristics that can be used for people classification based on appearance, including: physical, behavioral or adhered (such as clothing) features. Semantic segmentation of clothes is still a challenge for researchers because of the wide variety of clothing styles, layering, and shapes. This work presents an approach for clothing semantic segmentation tasks using the Feature Pyramid Network (FPN) with the EfficientNet as the backbone. We compare this approach with three other deep learning architectures: LinkNet, PSPNet, and U-Net. Due to the lack of a large dataset to train the deep learning model, we propose a combination of two datasets: CCP and CFPD, with refined labels to reduce similar classes. The resulting dataset contains 3,686 images with pixel-level annotations in 15 different categories. Experimental results show the effectiveness of our approach.**

***Index Terms*—Clothing Segmentation; Deep Learning; Clothing Parsing; Soft Biometrics**

I. Introduction

Human identification is a challenging problem in computer vision that has been studied over the past years. Robust human identification systems may provide real-time physical security, increase safety, and prevent crimes [1].

In the last years, the use of physiological or behavioral human characteristics has been used for the identification task. Human attributes, such as gender, age, hair color, clothes color, and tattoos are called soft biometrics and may allow distinguishing two individuals [2]. These attributes are often the first characteristics used when describing a person. In contrast to traditional approaches, soft biometrics extraction is a non-invasive and non-contact method that has the advantage of being performed at a distance without the cooperation of the targets.

Considerable efforts have been made into the problem of person re-identification using soft biometrics [3] . Most previous approaches have used demographics (age, gender, ethnicity, and race) and anthropometric (shape of the face, body, and skeleton) attributes [4]. Approaches using colors and types of clothes have received substantial attention lately due to the importance of these attributes in surveillance.

Recent advances in deep learning methods have achieved the state-of-art in many fields of study, including object classification and action recognition, and have achieved promising results in the segmentation task. Semantic segmentation using deep learning techniques is often addressed as a classification problem, and consists of assigning a semantic label to each pixel of an image [5].

Deep learning methods have also achieved promising results in the semantic segmentation of soft biometrics such as type and color of clothes, hair, and skin [6], [7]. These attributes are useful to many tasks, including identifying a person in surveillance videos or describe or semantically enrich a description of images. However, clothing segmentation is still a difficult task due to the wide similarity in different types of clothing or partial occlusion by a coat, scarf, hat, bag or other accessories. Fig. 1 depicts these problems by presenting different styles of garments such as dresses, coats, and pants.

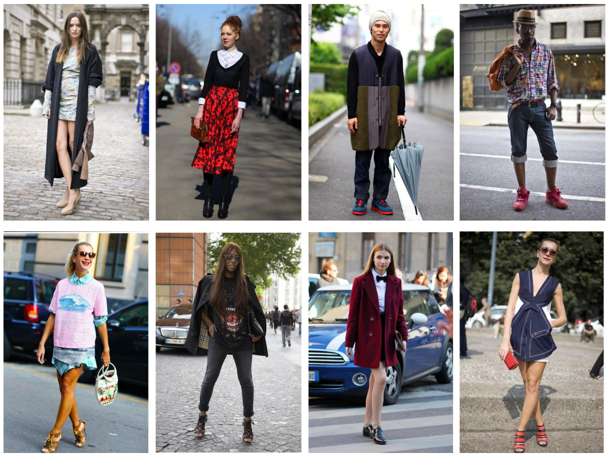

Fig. 1. Different colors and styles of clothes from the CCP dataset [8].

This work presents a deep learning-based approach to segmentation of clothes in the context of soft biometrics. We use the FPN Network [9] to perform this task. We also compare the obtained results with three other deep learning models models: Linknet [10], U-Net [11] and PSPNet [12]. To train these algorithms, we used two combined datasets: CCP [8] and CFPD [13] with pixel-level annotations. The resulting dataset contains a variety of clothing types and styles. The main contributions of our work are summarized as follow:

- Proposition of an approach for clothing segmentation task.
- Investigation and comparison of different deep learning architectures applied to the segmentation problem.
- Proposition of combining two datasets with the redefinition of labels to reduce similar classes.
- Provide a new benchmark with evaluation methodology for clothing segmentation task.

The rest of this paper is organized as follows. Section II gives a brief description of related works. Section III presents a thorough description of the Feature Pyramid Network (FPN) and the datasets used in this research. Section IV describes in detail the proposed method to extract soft biometric traits. Section V presents the experimental results and discussion. Finally, Section VI reports general conclusions and suggests future research directions.

II. Related Works

In the field of Computer Vision, several approaches have been proposed to classify, localize and segment clothes in several contexts,including soft biometrics, fashion analysis and surveillance.

Perlin and Lopes [2] proposed an approach to classify soft biometric traits using Convolutional Neural Networks (CNN). They used independent classifiers to detect the gender (Male / Female), Upper Clothes (short or long sleeves) and Lower Clothes (short or pants) of a person. Despite their approach achieved good generalization capability of the model, the authors reported difficulties in finding appropriate image datasets, concerning size, quality, and variability.

Hrkac, Brkic, and Kalafatic [14] presented a method for the segmentation of clothes using an adaptation of the U-Net deep learning architecture. This model was adapted to accommodate multi-class segmentation and was trained with the Clothing Co-Parsing (CCP) dataset. Due to the great similarity of classes with few instances, 58 different categories were grouped into 14. The authors conclude that the U-Net model can be a reliable way to perform the segmentation task and point out the lack of large datasets with pixel-level annotations.

Similar to [14] and [2], Tangseng, Wu and Yamaguchi [15] proposed a deep learning approach for clothing parsing tasks. Their architecture is based on Fully-Convolutional Neural Networks(FCN) and introduces a side path to FCN called "outfit encoder" to filter inappropriate clothing combination from segmentation, and a post-processing step using Conditional Random Field (CRF) to assign a visually consistent set of clothing labels. The authors used the CFPD and Fashionista dataset with refined labels to 25 categories to avoid similar labels.

The use of deep learning has demonstrated promising results in the clothing segmentation task. However, common challenging problems reported by authors are the lack of annotated data required by deep learning models, the use of dataset with unbalanced classes, and high similarity between categories. To overcome these problems, we propose to combine two datasets to increase the number of instances per class and use data augmentation techniques to increase the performance of the deep learning model. Besides that, we provide a benchmark for future comparison.

III. Background

A. Feature Pyramid Network (FPN)

Feature Pyramid Network (FPN) [9] is a general-purpose deep learning model that uses a top-down architecture with lateral connections to build high-level semantic feature maps in different scales using the image pyramid principle. This architecture has shown significant improvement as a generic feature extractor in many applications including object detection and instance object segmentation [16], [17].

FPN architecture, as presented in Fig. 2, consists of three components: a bottom-up pathway, a top-down pathway, and lateral connections.

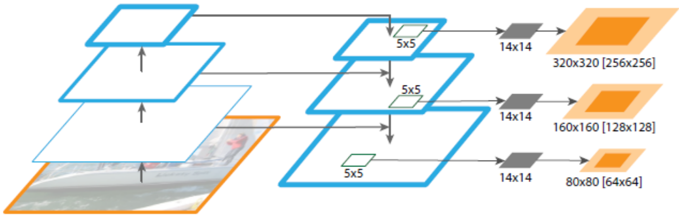

Fig. 2. FPN architecture for object segment proposal [9].

The bottom-up pathway is a traditional feed-forward (backbone) deep Convolutional Networks (ConvNet), such as ResNet [18] and VGG [19], which computes a feature hierarchy consisting of feature maps at several scales with a scaling step of 2. ConvNet has many layers producing output maps of the same size. These layers consist of a stage, and the last layer of each stage is used as the reference set of feature maps, which is enriched to create the pyramid.

The top-down pathway uses feature maps introduced in a pyramid during the bottom-up pathway and creates the final set of feature maps. The first layer of the top-down pathway receives from the last bottom-up layer the coarser-resolution feature map. Then, the spatial resolution is upsampled by factor 2 and merged with the corresponding bottom-up map by element-wise addition. Before the merge, a 1x1 convolutional layer is performed to reduce the channel dimension. The process of upsampling and merge is repeated until the last, and finest, resolution map is generated.

For segmentation purposes, a convolutional layer, with a kernel size of 3x3, followed by batch normalization and RELU activation, is applied twice successively. Then, these feature maps are upsampled to the same size and concatenated. Finally, the number of channels is decreased to the number of classes and upsampled to the original image size.

B. Dataset

The dataset employed in this work consists of a combination of two publicly and widely used datasets in clothing segmentation studies: the Clothing Co-Parsing (CCP) Dataset [8] and Colorful Fashion Parsing (CFPD) dataset [13].

The CCP dataset[1] contains a total of 2,098 high-resolution fashion photos, of which 1,004 with pixel-level annotations in 58 different labels plus background. The rest of the images are labeled only at the image-level, and we did not use it in this work. CFPD dataset[2] contains 2,682 images annotated with both colors (13 types) and categories (23 different types, including background). Fig. 3 shows some examples of CCP dataset images with pixel-level annotations.

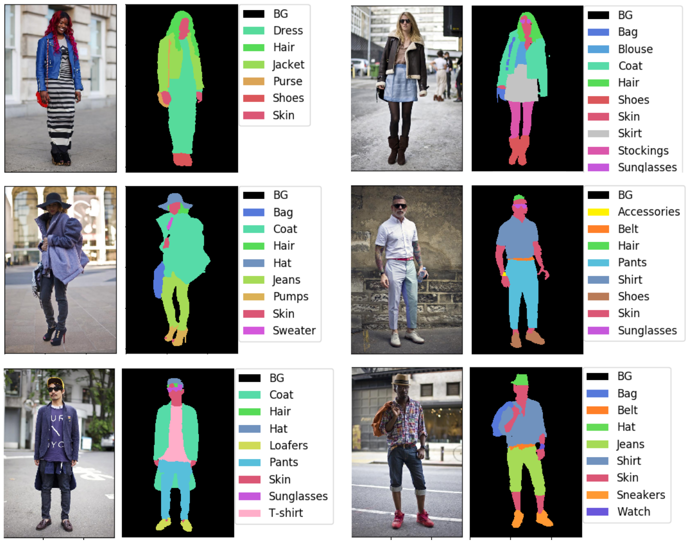

Fig. 3. Example of annotations in CCP [8] and CFPD [13] dataset .

In total, the combined dataset proposed in this work has 3,686 images with pixel-level annotations. Since the combined dataset is unbalanced and contains many similar classes with few examples, we redefined the labels, as presented in Table I. All small non-clothing objects presented in images, such as belt, bra, bracelet, and so on, were redefined to the background label. Classes without instances in the dataset were dismissed.

Nevertheless, the combined dataset is still unbalanced since the background and skin classes are present in all instances and fill a large part of the image. Fig. 4 presents the resulting class frequency distribution in the dataset.

IV. Methods

This work addresses the use of deep learning methods applied to the clothing segmentation problem. Fig. 5 presents an overview of the proposed approach, explained in the following sections.

A. Pre-processing

The first step consists in detecting, localizing, and cropping a person from an image. For this task, we use the Single Shot MultiBox Detector (SSD) [20], which uses Convolutional Networks to generate corresponding bounding boxes for each class instance. Then, cropped images are resized to 320x320, and each ground truth class is encoded using the one-hot encoding scheme. Finally, color images are normalized to the range 0 and 1.

Tabela I
Refined Labels

N.	Label	Original labels from CCP and CFPD dataset
1	BG	Acessories, Belt, BG, Bra, Bracelet,Earrings, Gloves, Necklace, Ring, Wallet, Watch
2	Bag	Bag, Purse
3	Coat	Blazer, Cape, Cardigan, Coat, Jacket, Suit
4	Dress	Dress, Skirt
5	Footwear	Boots, Clogs, Flats, Heels, Loafers, Pumps, Sandal, Shoes, Sneakers, Wedges
6	Glasses	Glasses, Sunglasses
7	Hair	Hair
8	Hat	Hat
9	Pants	Jeans, Leggings, Pants, Tights
10	Romper	BodySuit, Romper
11	Scarf/Tie	Scarf, Tie
12	Shirt	Blouse, Hoodie, Jumper, Shirt, Sweater, Sweatshirt, T-shirt, Top, Vest
13	Shorts	Shorts
14	Skin	Skin
15	Socks	Socks, Stockings

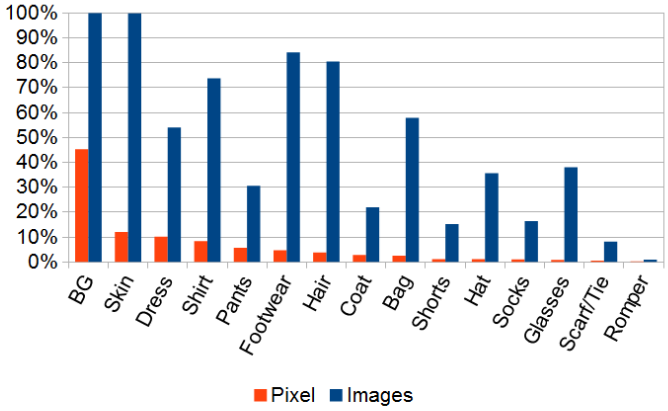

Fig. 4. Label statistics per pixel and images in the combined dataset.

B. Training

In this work, the clothing segmentation task was formulated as a classification problem. Therefore, a classifier is trained to classify each pixel into a target label. For this task, we use the Feature Pyramid Network (FPN). This architecture allows to use models with reliable capabilities as the encoder, and it's flipped version as the decoder. The pre-trained weights were obtained from networks trained on the 2012ILSVRC ImageNet dataset [21]. We also compare the results obtained with three deep learning models for segmentation tasks: LinkNet [10], Pyramid Scene Parsing Network (PSPNet) [12] and U-Net [11]. The EfficientNet [22] model was used as a backbone network in all tested architectures, which has got the state-of-the-art performance in ImageNet dataset recently.

[1] http://www.sysu-hcp.net/clothing-co-parsing-by-joint-image-segmentation-and-labeling/

[2] https://github.com/hrsma2i/dataset-CFPD

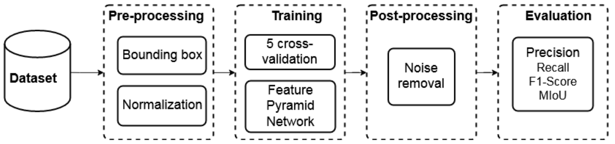

Fig. 5. Overview of the proposed method.

Deep learning methods require massive volumes of data for training. However, due to the lack of large training and testing sets, online data augmentation techniques including flip, rotation, random crop, Gaussian noise, and variation of brightness and contrast were used in the training set to improve the generalization capability of the model.

The training process was performed for 100 epochs or until the accuracy stagnates for ten consecutive epochs. The Adam optimizer was used with a learning rate of 0.001. A learning rate reduction by a factor of 0.1 was employed when the validation loss metric has stopped improving to improve the overall performance during the training process.

The 5-fold cross-validation method was conducted to evaluate the model generalization performance. The stratified sampling method, proposed by [23], was applied to ensure the same class distribution in each generated subset. All folds were divided into a ratio of 80/20.

C. Post-processing

The output of the training model consists of a softmax probability map of the same size of the image with the number of channels defined by the number of classes. The opening morphological operation is applied to each channel of the predicted output to reduce predicted noises. Then, the binary feature maps are merged using the argmax operation to obtain the final prediction mask.

D. Evaluation

The Precision, Recall, F1-score and Intersection over Union (IoU) metrics, given by Equations 1, 2, 3 and 4, respectively, are commonly evaluation metrics score used for clothing semantic segmentation [24] and were used to analyze the outcome of the experiments.

Precision is a measure that indicates the proportion of predictions that are true positives. Recall is a measure of completeness and specifies the proportion of positives that are detected: F1-score combines Precision and Recall and is the harmonic mean of these two measures. IoU refers to the intersection of the ground truth and the predicted segmentation divided by the union of the ground truth and the predicted segmentation.

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i}, \tag{1}$$

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i}, \tag{2}$$

$$\text{F1-Score}_i = 2\frac{Precision * Recall}{Precision + Recall}, \tag{3}$$

$$\text{IoU} = \frac{p_{ii}}{\sum_{j=1}^{nc}(p_{ij}) + \sum_{j=1}^{nc}(p_{ji}) - p_{ii}}. \tag{4}$$

Where TP indicates the True Positive, FP indicates the False Positive and FN indicates the False Negative for a class i.

We did not report accuracy since it is not a good metric to measure performance on an unbalanced dataset, as described by [25].

V. Experimental Results and Discussions

This Section presents experimental results obtained by the segmentation using the method described in Section IV. Experiments were performed on a workstation with IntelCore-i7 8700 processor, 32GBytes RAM, and a Nvidia Titan Xp GPU. The Keras library using the TensorFlow backend was used to train and test the models.

We compare four deep learning approaches to the clothing segmentation task. The 5-fold cross-validation method was used to assess the generalization performance of models over the dataset. By the end of each fold, the best model was used to predict the samples in the test set and measure the performance using metrics described in Section IV-D. Results were evaluated quantitatively, calculated from the average of the metrics obtained by each fold, and qualitatively, by visual inspection of the segmented images.

A. Quantitative results

Table II shows the performance of the four deep learning models evaluated in this work. We notice that the FPN model achieved the best results in terms of mean Precision, Recall, and F1-Score, indicating that this model performs semantic segmentation better than other models in the dataset used in this work.

Tabela II
Comparison of the Precision, Recall and F1-Score metrics on the test set with four deep learning models.

N.	Model	Precision	Recall	F1-Score
1	LinkNet	74.8%	77.2%	74.9%
2	PSPNet	71.9%	69.7%	70.3%
3	U-Net	75.3%	78.2%	76.7%
4	FPN	**76.6%**	**80.4%**	**77.9%**

Table III shows the clothing segmentation performance of each class predicted by the FPN model. We can notice that background, pants, dress, and skin classes achieve the best results in terms of F1-Score. On the other hand, romper, hat, and glasses achieve the poorest result. The low score achieved by these classes can be related to the small number of examples available in the dataset.

The romper class has only 34 instances, or 0.009% of the dataset, as depicted in Fig. 4. Furthermore, it was also noted that this class is easily confused with Pants or Dress. Hat and Glasses classes also achieve poor results because they are small objects and therefore they have few pixels in the image. These results highlight the difficulty of achieving good results in less frequent classes.

Tabela III
PER-CLASS SEGMENTATION PERFORMANCE OBTAINED BY FPN MODEL.

N.	Class	Precision	Recall	F1-Score
1	Background	96.4%	93.2%	94.7%
2	Bag	82.1%	84.8%	83.4%
3	Coat	77.8%	85.9%	81.7%
4	Dress	89.8%	89.5%	89.7%
5	Footwear	77.3%	85.4%	81.1%
6	Glasses	46.6%	78.3%	58.4%
7	Hair	83.9%	84.6%	84.2%
8	Hat	61.6%	69.0%	65.1%
9	Pants	90.4%	90.6%	90.5%
10	Romper	53.4%	30.9%	39.2%
11	Scarf/Tie	68.6%	73.7%	71.1%
12	Shirt	85.1%	85.0%	85.0%
13	Shorts	72.6%	82.3%	81.3%
14	Skin	85.7%	88.2%	86.9%
15	Socks	78.0%	74.6%	76.2%

We also evaluate the detection performance of the model, reported in Fig. 6, by using the IoU metric. The Mean Intersection over Union, that consists of the average over all classes, achieves 65.7%. Most classes achieved adequate results, indicating that the proposed approach is reliable for the clothing segmentation task. According to [26], predictions with intersection over union more than 50% are considered satisfactory.

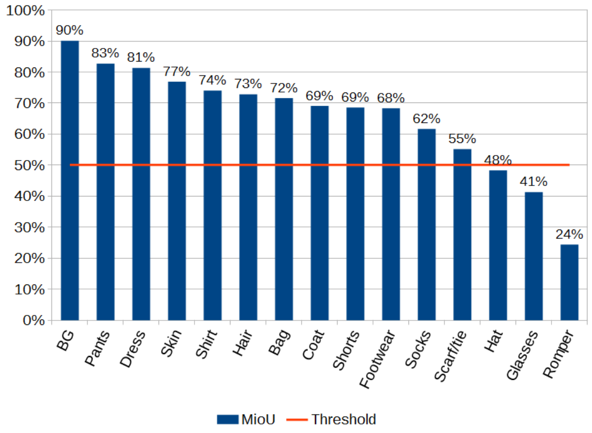

Fig. 6. Comparison of IOU scores for each category.

The confusion Matrix, presented in Fig. 7, shows the predicted and target classifications. It is observed that most classes have the highest values distributed on the main diagonal, attesting the effectiveness of our approach.

Small objects, such as glasses and socks, have achieved satisfactory results, despite having some classification errors. For instance, socks occasionally were wrongly predicted as pants or background and glasses were predicted as hair or skin.

This suggests that more data is required to improve the classification results. However, by increasing the number of instances with these classes, consequently, the number of pixels in other classes such as skin and hair also increases. This situation makes the clothing semantic segmentation task even more challenging.

We may also note that many rompers were predicted as dress, pants, and shirt. This highlights the difficulty in the task of clothes segmentation, as some classes are very similar among themselves.

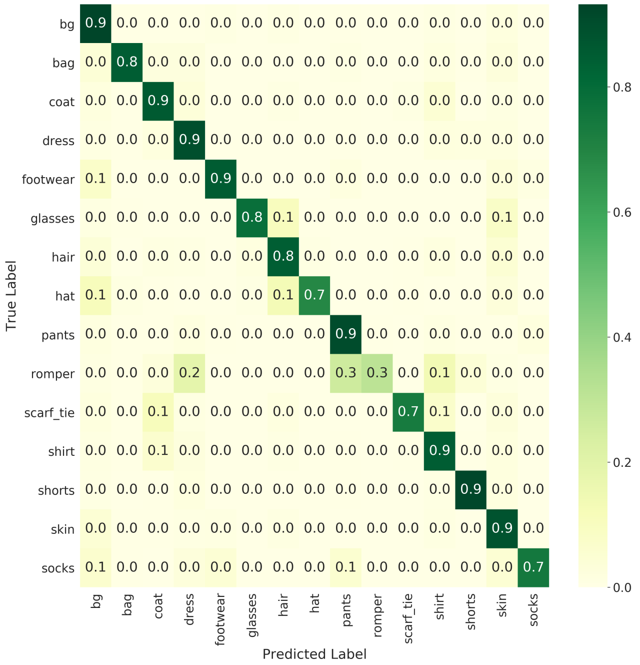
	bg	bag	coat	dress	footwear	glasses	hair	hat	pants	romper	scarf_tie	shirt	shorts	skin	socks
bg	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
bag	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
coat	0.0	0.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
dress	0.0	0.0	0.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
footwear	0.1	0.0	0.0	0.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
glasses	0.0	0.0	0.0	0.0	0.0	0.8	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
hair	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
hat	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
pants	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0
romper	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.3	0.3	0.0	0.1	0.0	0.0	0.0
scarf_tie	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.1	0.0	0.0	0.0
shirt	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9	0.0	0.0	0.0
shorts	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9	0.0	0.0
skin	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9	0.0
socks	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.7

Fig. 7. The confusion matrix representing the computed pixel classification accuracy %.

B. *Qualitative results*

In this section, we present a visual evaluation of the outputs predicted by the proposed approach. Fig. 9a) presents sample images of the testing set with ground truth and output provided by the FPN network. The trained model was able to segment different types of clothes satisfactorily and confirm that our approach is robust enough to this task.

Poor segmentation results were also predicted by our model and it is presented in Fig. 9b). Even after redefining labels to avoid similar classes, as described in Section III-B, we still note similar classes that caused misclassifications. For instance, the socks class is occasionally confused with pants or skin.

Rompers also achieved unsatisfactory segmentation results. This class consists of a one-piece or two-piece combination of shorts or pants and a shirt or blouse with short or long sleeves on the top. Moreover, romper was usually classified as pants or dresses due to the similarity in shape among these classes, as presented in Fig. 8.

It is possible to observe that romper has different shapes and styles, and the classifier did not learn satisfactorily how to classify this class. This may indicate that the model requires more training data or these classes are not easily separable.

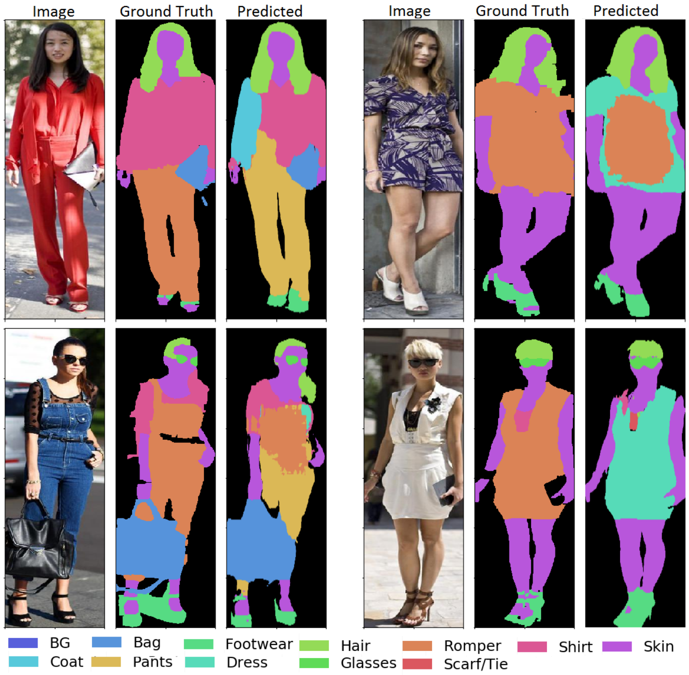

Fig. 8. Parsing results of the Rompers category.

A few annotation errors were observed on both datasets, which require improvements. It was also noted a few images in both datasets with background annotation only. This condition is not desired since it may cause prediction errors.

VI. Conclusions

Image segmentation has been one of the most difficult problems in computer vision that could be used to improve applications in many areas including security and surveillance. Recently, soft biometrics traits, including types of clothes, have been shown promising results in person re-identification. However, it is still a difficult task to solve due to the high variety of types, styles, shapes, and layering.

Although semantic segmentation using deep learning algorithms has achieved great success in many research fields, it still challenging for computers to understand and describe a scene as humans do.

In this paper, we propose an approach based on the Feature Pyramid Network (FPN) for clothing semantic segmentation in the context of soft biometrics. A comparison with three deep learning methods was performed. These architectures have been proposed recently and achieved state-of-the-art performance on various datasets.

Due to the lack of large datasets with pixel-level annotations, we combine the two public datasets, CCP and CFPD, resulting in more representative training data and improving model generalization.

Quantitative and qualitative results presented show the effectiveness of the approach. Results obtained in this work have the potential to improve many real applications such as clothes recommendation, person re-identification, image retrieval, image description, and surveillance.

Future works include testing other deep learning architectures, improve the annotation quality of both datasets, explore new methods for improving the segmentation of similar objects and extract colors from segmented objects.

Acknowledgment

A. S. Inácio thanks for support from the Federal Institute of Santa Catarina; A. Brilhador thanks for support from the Federal Technological University of Paraná; H.S. Lopes thanks CNPq for the research grants no. 311778/2016-0 and 423872/2016-8, and Fundação Araucária for the research grant PRONEX 042/2018. Special thanks to NVIDIA Corp. for the donation of the Titan-Xp GPU boards used in this work.

Bibliografia

[1] A. Costin, "Security of cctv and video surveillance systems: Threats, vulnerabilities, attacks, and mitigations," in *Proceedings of the 6th International Workshop on Trustworthy Embedded Devices*. New York, NY: ACM, 2016, pp. 45–54.

[2] N. Romero, M. Gutoski, L. Hattori, and H. Lopes, "Soft biometrics classification using denoising convolutional autoencoders and support vector machines," in *Anais do 13 Congresso Brasileiro de Inteligência Computacional*. Curitiba, PR: ABRICOM, 2017, pp. 1–12.

[3] A. Abdelwhab and S. Viriri, "A survey on soft biometrics for human identification," in *Machine Learning and Biometrics*, J. Yang, D. S. Park, S. Yoon, Y. Chen, and C. Zhang, Eds. Rijeka: IntechOpen, 2018, ch. 3.

[4] A. Dantcheva, P. Elia, and A. Ross, "What else does your biometric data reveal? a survey on soft biometrics," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 3, pp. 441–467, 2016.

[5] K. Yue, L. Yang, R. Li, W. Hu, F. Zhang, and W. Li, "Treeunet: Adaptive tree convolutional neural networks for subdecimeter aerial image segmentation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 156, pp. 1 – 13, 2019.

[6] Z. Chen, S. Liu, Y. Zhai, J. Lin, X. Cao, and L. Yang, "Human parsing by weak structural label," *Multimedia Tools and Applications*, vol. 77, no. 15, pp. 19 795–19 809, 2018.

[7] P. Tangseng, Z. Wu, and K. Yamaguchi, "Looking at outfit to parse clothing," *CoRR*, vol. abs/1703.01386, 2017.

[8] W. Yang, P. Luo, and L. Lin, "Clothing co-parsing by joint image segmentation and labeling," in *Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, 2013, pp. 3182–3189.

[9] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, 2017, pp. 2117–2125.

[10] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," in *Proceedings of IEEE Visual Communications and Image Processing (VCIP)*, St. Petersburg, FL, 2017, pp. 1–4.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, vol. 9349. Munich: Springer International Publishing, 2015, pp. 234–241.

[12] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, 2017, pp. 2881–2890.

[13] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, and S. Yan, "Fashion parsing with weak color-category labels," *IEEE Transactions on Multimedia*, vol. 16, no. 1, pp. 253–265, 2014.

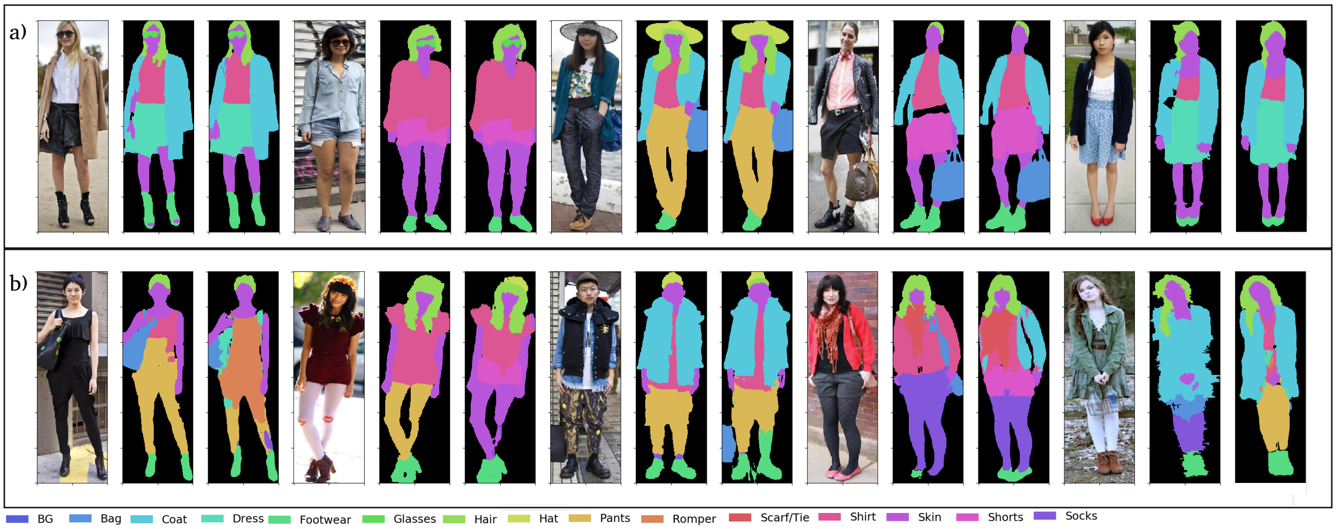

Fig. 9. Parsing results on the test set based on the a)highest and b)lowest IOU score. The figure shows the input image, ground truth, and the predicted from left to right, respectively.

[14] T. Hrkac, K. Brkic, and Z. Kalafatic, "Multi-class u-net for segmentation of non-biometric identifiers," in *Proceedings of the 19th Irish Machine Vision and Image Processing conference*, Dublin, 2017, pp. 131 – 138.

[15] H. A. Perlin and H. S. Lopes, "Extracting human attributes using a convolutional neural network approach," *Pattern Recognition Letters*, vol. 68, pp. 250–259, 2015, special Issue on "Soft Biometrics.

[16] S. Seferbekov, V. Iglovikov, A. Buslaev, and A. Shvets, "Feature pyramid network for multi-class land segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Salt Lake City, UT, 2018, pp. 272–275.

[17] J. Ji, S. Buch, A. Soto, and J. C. Niebles, "End-to-end joint semantic segmentation of actors and actions in video," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, 2018, pp. 702–717.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 770–778.

[19] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," in *Proceedings of 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, Kuala Lumpur, 2015, pp. 730–734.

[20] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Proceedings of the 14th European Conference on Computer Vision*. Amsterdam: Springer International Publishing, 2016, pp. 21–37.

[21] P. Yakubovskiy, "Segmentation models," 2019. [Online]. Available: https://github.com/qubvel/segmentation_models

[22] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 97, Long Beach, CA, 2019, pp. 6105–6114.

[23] K. Sechidis, G. Tsoumakas, and I. Vlahavas, "On the stratification of multi-label data," in *Proceedings of the 2011 European conference on Machine learning and knowledge discovery in databases*, Berlin, Heidelberg, 2011, pp. 145–158.

[24] F. Lateef and Y. Ruichek, "Survey on semantic segmentation using deep learning techniques," *Neurocomputing*, vol. 338, pp. 321 – 348, 2019.

[25] G. E. A. P. A. Batista, A. C. P. L. F. Carvalho, and M. C. Monard, "Applying one-sided selection to unbalanced datasets," in *Proceedings of Mexican International Conference on Artificial Intelligence*, Berlin, Heidelberg, 2000, pp. 315–325.

[26] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015.