Real-Time Gender Detection in the Wild Using Deep Neural Networks

Luis Felipe Zeni
Informatics Institute
Federal University of Rio Grande do Sul
Porto Alegre, Brazil
Email: luis.zeni@inf.ufrgs.br

Claudio Jung
Institute of Informatics
Federal University of Rio Grande do Sul
Porto Alegre, Brazil
Email: crjung@inf.ufrgs.br

***Abstract*—Gender recognition can be used in many applications, such as video surveillance, human-computer interaction and customized advertisement. Current state-of-the-art gender recognition methods are detector-dependent or region-dependent, focusing mostly on facial features (a face detector is typically required). These limitations do not allow an end-to-end training pipeline, and many features used in the detection phase must be re-learned in the classification step. Furthermore, the use of facial features limits the application of such methods in the wild, where the face might not be present. This paper presents a real-time end-to-end gender detector based on deep neural networks. The proposed method detects and recognizes the gender of persons in the wild, meaning in images with a high variability in pose, illumination an occlusions. To train and evaluate the results a new annotation set of Pascal VOC 2007 and CelebA were created. Our experimental results indicate that combining both datasets during training can increase the mAp of our gender detector. We also visually analyze which parts leads our network to make mistakes and the bias introduced by the training data.**

I. Introduction

Object detection methods presented a significant increment in accuracy and velocity in recent years [1], [2]. These gains are mostly related to: (i) the availability of larger datasets; (ii) the parallelism allowed by the usage of GPUs; and (iii) the use of deep architectures such as Convolutional Neural Networks (CNNs) [3]. Object detectors are build to deal with a set of different classes of objects. However, in many computer vision applications (i.e., pedestrian counting, fruit selection, car license plate localization) there is no need to detect a large number of classes. Other applications need to deal with intra-class fine granularity detection, for example, to detect different car models, determine the breed of a cow or a person gender.

Gender information is a relevant feature in many computer vision applications such as video surveillance, human-computer interaction, statistics about consumer preferences, and gender oriented advertisement. Most gender detection methods in the literature are based exclusively in facial cues, since the human face is a good source of discriminative gender features. As an example, the approach presented in [4] achieves 94% of accuracy using facial features in the LFWA dataset [5].

Despite the high accuracy obtained by recent facial-based gender recognition approaches, they tend to fail when the face presents low resolution, is partially occluded or not event present in the image. Nowadays, millions of images acquired in a wide variety of situations and scenarios are daily uploaded into social media networks by its users, which are acquired in a wide variation of the person pose, body/face occlusions, deformations and illumination changes. As a motivational example, a company that desires to crawl images from social media to evaluate the gender of people that appear associated with the company logo would encounter several challenges if using a face-based method, as illustrated in Figure 1. In this case, face-based methods will only recognize the male on the left, failing to detect/recognize the two other people in the scene. As in other computer vision problems, we refer to gender detection with less constraints as *gender detection in the wild*.

Fig. 1. Example of gender detection in the wild using the proposed method. The bounding box color indicates the gender, along with the confidence value of the winner class (i.e the class with the highest confidence).

There are also a few methods that explore other features present in the human body in the context of gender recognition. This class of methods explores images of pedestrians, which usually are in low resolution and in common poses such as walking or standing. These limitations lead to low accuracy (for example, Tian et al. [6] report a Log-average miss rate of 30.70% in the PETA dataset [7]).

Another limitation of most existing methods is that they are detector-dependent [3], [6], [8], [9] or region-proposal-dependent [4]. In other words, these methods rely on external modules, leading to a complex training pipeline. Furthermore, this dependency does not allow the sharing of features between detection and recognition, turning the whole process slower and not getting advantage of an end-to-end training process.

Although CNN-based models have presented unprecedented results in the last years, understanding and explaining which regions of a given image guides a CNN to its result is crucial, mainly in cases of classification/detection mistakes [10]. Such localized analysis can help us understand why a network succeeds, misses or explores contextual information.

In this paper we propose a novel end-to-end real-time gender detection/recognition method in the wild using Regional Convolutional Neural Networks. As (to best of our knowledge) there are no available annotated datasets to this new task, an additional contribution is a new set of gender-related annotations to the Pascal VOC 2007 [11] and CelebA datasets [12][1]. We also evaluate the effect of using different proportions of each dataset during the training process and compare our detector with other gender recognition methods. Finally, we present a visual analysis of the regions that contribute to successes/misses and the developed biases of the proposed gender detector method.

The remainder of this paper is organized as follows. Section II revises some related gender recognition approaches, datasets, and activation visualization methods. Section III describes the proposed gender detection method in detail. Section V presents the experimental results and discussions. Section VI provides the conclusions.

II. Related Work

Before describing the proposed method, we briefly review related methods and datasets for gender recognition, focusing on CNN-based methods. Previous CNN methods use one or more combination of features such as LBP, SURF, HOG or SIFT [4], and a detailed survey of gender recognition methods can be found in [13], [14]. We also review methods for visualizing neuronal activation, which is useful for understanding the results of a given CNN model in an image.

A. Gender Recognition

We separated the reviewed methods into two classes: (i) methods based only in facial features; and (ii) methods based on body features.

Face Based. This class of methods focuses only on facial features for gender classification. Although the face indeed presents relevant gender-related features, it is frequently occluded or not present in the context of images in the wild. Some existing methods also extract other facial attributes such as age and pose in conjunct with gender [3], [4].

In [3] a CNN-based method is proposed to classify gender and age, presenting only three convolution layers and using ReLU as the activation function. Dropout is used when training the model to avoid overfitting, and evaluation is performed using the Adience benchmarks dataset [15], which contains images in extreme blur, facial occlusions, pose variations and different facial expressions. The method relies on an external face detector to work, not allowing end-to-end training and sharing the features between detection and classification. In [8] a similar detector-dependent approach is proposed. A pre-trained CNN is fine-tuned using dropout to extract features from face images, and the feature vector is used as input to an SVM classifier that determines the gender class. This approach creates a complex training pipeline because the CNN and the SVM present distinct training pipelines. Therefore, the method is not able to optimize the feature extraction and the classifier together. The evaluation is done using the Adience dataset and also the FERET [16] dataset, which is smaller and less challenging.

Joint face detection and gender classification reduces computational cost and allows jointly training/fine-tuning for both tasks. Ranjan and colleagues [4] proposed an architecture that performs face detection and gender classification using CNNs, also estimating facial landmark points and pose. The authors indicate that learning different tasks in conjunction improve results of specific tasks. The Alexnet architecture [17] is used as the basis, and a set of extra convolution layers are connected into the network architecture. These convolution layers feed an intermediary general feature fully-connected layer, which is used as input to a set of different groups of fully-connected layers, each one specialized to detect a different attribute in a detected face. The detection process is similar to R-CNN [18], where region proposals are generated using Selective Search [19], and each region is independently classified by the network. However, feeding each region into the network is very expensive, turning the detection process slow [20]. The method uses the Annotated Facial Landmarks in the Wild (AFLW) dataset [21], which contains ∼25k faces in real-world images with pose, expression, ethnicity, age and gender variations.

Body Based. Recognizing gender using body features is a more complex task, as the body can present different configurations, occlusions, deformations and pose variations. Furthermore, males and females can be distinguished more easily using facial than body information. The reviewed methods detect features in the context of pedestrian detection. In [6], a pedestrian detector based on CNNs is proposed. The method jointly optimizes pedestrian detection with auxiliary semantic tasks, including pedestrian attributes (e.g. 'backpack', 'gender', and 'views') and scene attributes (e.g. 'vehicle', 'tree', and 'vertical'). The proposed network presents an architecture similar to AlexNet, with the addition of fully-connected layers at the end, each one specialized in one kind of attribute. The network was trained an evaluated in Caltech [7] dataset, which contains a total of 350,000 bounding boxes annotated with 2,300 unique pedestrians.

In [9], a Joint Recurrent Learning model is formulated for exploring attribute context and correlation to improve at-

[1]The annotation data and source code are publicly available at http://www.inf.ufrgs.br/~crjung/gender_sib2018

tribute recognition. The model uses Long-Short Term memory (LSTM) to learn jointly pedestrian attribute correlations in a pedestrian image, and in particular their sequential ordering dependencies. The key idea of the method is to encode sequentially localized person spatial contexts, and to propagate inter-region contextual information. The method uses the PETA dataset [22], which contains 19,000 images of 8,705 pedestrians annotated with different attributes.

B. Interpreting a learned CNN

With the remarkable results achieved by CNN-based methods in last years, some researchers directed attention to develop techniques that help to understand and try to explain the behavior of CNNs. Some methods aim to visualize the features and kernels learned by CNNs, as well as which regions in the image were "activated".

In [23], a deconvolution approach that uses guided Back-propagation was proposed. Zeiler and Fergus [24] proposed modifications into gradients that lead to improvements in the qualitative results. Even if these methods produce fine-grained visualizations, the results are not class-discriminative and the visualizations of different classes are nearly identical [10].

Other methods generate images that maximize the activations of a network unit [25] or invert a latent representation [26]. Although these methods synthesize high-resolution images with class-discriminative, they visualize a model globally and not specific instances of image predictions.

In contrast, methods based on gradient-weighted class activation mapping (Grad-CAM) use the gradients of any target concept flowing the final convolutional layer to produce a coarse localization map that highlights the importance of regions of a given image to the result. Zhou and colleagues [27] modified the CNN architecture by replacing fully-connected layers with convolutional layers and global average pooling. In [28], a similar method using global max pooling is investigated, while log-sum-exp pooling is explored in [29]. Recently, Selvaraju et al. [10] combine Grad-CAM with existing fine-grained visualizations to create a high-resolution class-discriminative visualization.

Although Grad-CAM methods can create a class-discriminative visualization, it takes all instances of a given class into account. This is a deficiency in the context of object detection, where it is important to know why a specific instance of the class was correctly detected and another not. In this paper we adapt Grad-CAM to work with R-CNNs, allowing the visualization of which pixels contributed more effectively towards a specific detection.

III. Real-Time Gender Detection in the Wild

In this work we present an approach for gender detection "in the wild", meaning that it expected to detect a `man` or `woman` in a variety of poses, occlusions and illumination conditions. We set the following goals for our method: (i) detect gender directly in images without depending on external modules (e.g., a person detector); (ii) work in real-time (24 FPS or more); (iii) use any information available in the image that is related to gender.

To our knowledge our work is the first to tackle gender detection/recognition in this way. Our first difficulty was to find a proper dataset with the needed characteristics for this work. Some datasets provide the images already cropped (i.e., only the face or the pedestrian is present) [7], [16], [21], [22]. Other datasets provide only facial bounding box annotated with gender [12], and in the case of the IMDB/Wiki dataset [30], the face bound box and the gender annotation are not reliable because the dataset was automatically crawled from the Internet. A limitation of these datasets is that they have only one instance annotated per image. Other popular datasets [11], [31] provide labeling of all present information of the class `person` in each image and contain more than one instance per image, but unfortunately the gender data is not provided with these datasets. To overpass the "data problem" with a low work cost, we adapted two existing databases to this problem. More precisely, we adapted the Pascal VOC2007 [11] and CelebA [12] datasets. Next, we describe what and how we adapted in each dataset.

Pascal VOC2007 adapted for gender detection. The Pascal VOC2007 dataset [11] contains 20 classes of objects, including people captured at different poses and occlusions. Many images present more than one instance of persons, and it is common to have intersections between bounding boxes. To adapt this dataset, we first selected only images and annotations containing persons. As the dataset already presents the annotated bounding boxes, we just changed the labels from `person` to `man`, `woman` or `undefined`. The undefined label is assigned to people for which the human annotator cannot estimate the gender. We also maintain the original "difficult" flag to all annotations and all undefined bounding boxes are also marked as "difficult". The manual labeling work was made by a single annotator, resulting in 4,192 (2,095 to train and 2,097 to test) annotated images with 4,381 instances of men, 3,210 instances of women and 3,083 instances of persons with undefined gender. This new annotation set presents a very challenging dataset because the images usually contain a high variation in pose, occlusion, intersections and multiple instances in one image. However, it contains too few images, and we also adapt the CelebA dataset to alleviate this limitation.

CelebA adapted for gender detection in the wild. The CelebA dataset [12] contains images of celebrities annotated with 40 attributes, including gender. The dataset contains one annotated person per image, and only the face bounding box is provided. To adapt this dataset to our task, we first run a person detector [1] in all CelebA images. From these detections, we selected the new ground truth bounding box, which is the detected bounding box with the largest intersection with the annotated face ground truth and the largest probability of being a person. This automatic process leads to 202,517 (182,562 to train and 19,955 to test) annotated images with 84,380 instances of men and 118,137 instances of women.

The Chosen Network Architecture Our method is inspired

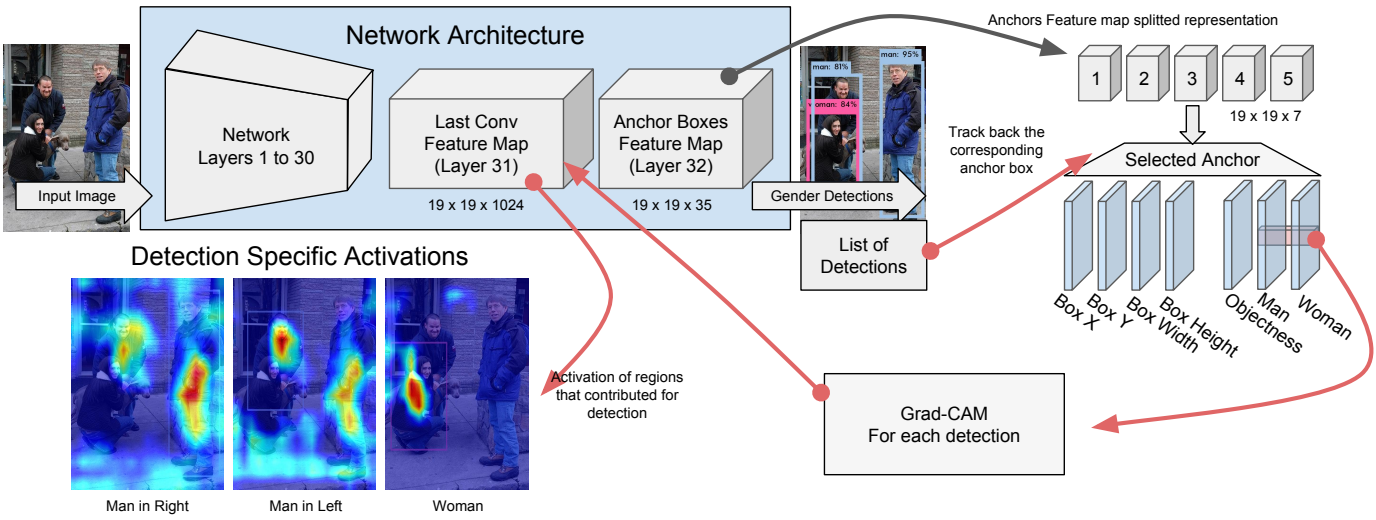

Fig. 2. Proposed Instance class discriminative Grad-CAM for R-CNN. First, the image is fed into the network to get valid detections. For each detection, we track it back to the anchor box that generates it with the corespondent x and y positions in the final layer. Then we use [10] with the respective x and y for the respective layers from classes 'man' and 'woman' to create the activation maps.

by the recent advances in object detection [1], [2]. To avoid the dependency of an external region proposal method, we use a region proposal network (RPN) similar to the one proposed in [32], which predicts offsets and confidences for anchor boxes in the final layer of the network. Since the prediction layer is convolutional, the RPN predicts these offsets at every location in a feature map [1]. Our model architecture is based on the YOLO v2 [1] architecture with 31 layers, as illustrated in Table I. The anchors layer had the size reduced to $19 \times 19 \times 35$ to work only with two classes, similarly to other works that customized YOLO for a two-class problem [33], [34]. To avoid having similar bounding boxes with different labels, we also used non-maxima suppression (NMS) as a post-processing step. More precisely, if two detection results with different labels present an intersection over union (IoU) larger than a threshold T_{IoU} (set to 0.4 experimentally), only the class with the highest confidence is retrieved.

Transfer Learning and Data Augmentation. A common practice in deep learning is to use "transfer learning" to avoid overfitting and to accelerate the training process. Basically, this technique initializes the new network with the weights of a pre-trained network and performs fine-tuning using the target dataset. We initialize the weights of layers 1 to 30 in our model with a pre-trained version of Yolo V2 based on the Imagenet and coco datasets, and the last two layers were initialized with random values. We then fine-tune the weights of all layers toward gender detection/recognition using the CelebA and Pascal VOC datasets. We also used data augmentation and batch normalization during the training process, which leads to significant improvements in convergence while eliminating the need for other forms of regularization like dropout [1].

TABLE I
GENDER DETECTION NETWORK: AN ADAPTATION OF THE YOLO NETWORK TO 2 OBJECT CLASSES (MAN/WOMAN).

#	Layer	Filters	Size	Input	Output
1	conv	32	3 x 3 / 1	608 x 608 x 3	608 x 608 x 32
2	max		2 x 2 / 2	608 x 608 x 32	304 x 304 x 32
3	conv	64	3 x 3 / 1	304 x 304 x 32	304 x 304 x 64
4	max		2 x 2 / 2	304 x 304 x 64	152 x 152 x 64
5	conv	128	3 x 3 / 1	152 x 152 x 64	152 x 152 x 128
6	conv	64	1 x 1 / 1	152 x 152 x 128	152 x 152 x 64
7	conv	128	3 x 3 / 1	152 x 152 x 64	152 x 152 x 128
8	max		2 x 2 / 2	152 x 152 x 128	76 x 76 x 128
9	conv	256	3 x 3 / 1	76 x 76 x 128	76 x 76 x 256
10	conv	128	1 x 1 / 1	76 x 76 x 256	76 x 76 x 128
11	conv	256	3 x 3 / 1	76 x 76 x 128	76 x 76 x 256
12	max		2 x 2 / 2	76 x 76 x 256	38 x 38 x 256
13	conv	512	3 x 3 / 1	38 x 38 x 256	38 x 38 x 512
14	conv	256	1 x 1 / 1	38 x 38 x 512	38 x 38 x 256
15	conv	512	3 x 3 / 1	38 x 38 x 256	38 x 38 x 512
16	conv	256	1 x 1 / 1	38 x 38 x 512	38 x 38 x 256
17	conv	512	3 x 3 / 1	38 x 38 x 256	38 x 38 x 512
18	max		2 x 2 / 2	38 x 38 x 512	19 x 19 x 512
19	conv	1024	3 x 3 / 1	19 x 19 x 512	19 x 19 x 1024
20	conv	512	1 x 1 / 1	19 x 19 x 1024	19 x 19 x 512
21	conv	1024	3 x 3 / 1	19 x 19 x 512	19 x 19 x 1024
22	conv	512	1 x 1 / 1	19 x 19 x 1024	19 x 19 x 512
23	conv	1024	3 x 3 / 1	19 x 19 x 512	19 x 19 x 1024
24	conv	1024	3 x 3 / 1	19 x 19 x 1024	19 x 19 x 1024
25	conv	1024	3 x 3 / 1	19 x 19 x 1024	19 x 19 x 1024
26	route	16			
27	conv	64	1 x 1 / 1	38 x 38 x 512	38 x 38 x 64
28	reorg		/ 2	38 x 38 x 64	19 x 19 x 256
29	route	27 24			
30	conv	1024	3 x 3 / 1	19 x 19 x 1280	19 x 19 x 1024
31	conv	35	1 x 1 / 1	19 x 19 x 1024	19 x 19 x 35
32	detection				

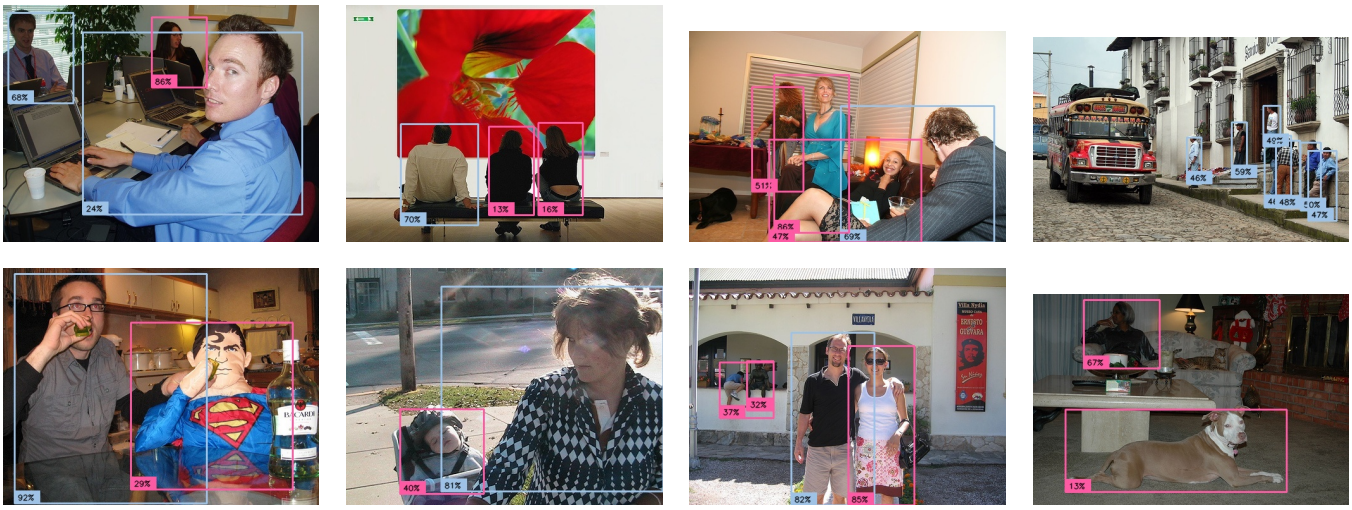

Fig. 3. First row presents visual examples of successes and the second errors of the proposed method (Using the 50%+50% model and $t = 0.1$).

IV. Instance class discriminative Grad-CAM for R-CNN

A CNN learns a hierarchy of patterns and concepts to represent the visual world using convolutional filters. These filters extract features that retain spatial information, which is lost in fully-connected layers. Following this hierarchical concept, the last convolution layers (neurons) have the best compromise between high-level semantics class-specific information and detailed spatial information [10], [24], [26]. To understand the importance of each of these neurons, Grad-CAM methods operate in the gradient information from the fully-connected layer in the direction of the last convolutional layer in the CNN.

We adapted the method proposed by Selvaraju et al. [10] to visualize the activations of detections produced by R-CNN networks. Their work allows to visualize activated regions that contributed to a specific class in a given image, but the resultant activation map of a class reacts to all instances present in the image. For object detection, this visualization will not help much, as it does not explain which regions contribute to the decision process for each detected object.

Our approach uses the final anchors feature map to create the visualization of regions that contribute to each detection. Each detection is tracked back to the corresponding anchor box φ_{det} and the spatial locations $x, y \in \{1, ..., 19\}$ in the feature map, which presents a spatial resolution of 19×19 in the detection layer.

Finally, we backpropagate the information from the point (x, y) in anchor φ_{det} and layer corresponding to either `man` or `woman` (the detected class) along the network to produce a Grad-CAM localization map for this detection. We used the same Grad-CAM framework proposed in [10] to create the activation maps. The pipeline of the proposed visualization method is presented in Figure 2.

V. Experiments

We evaluated our method by training different models that "blend" different proportions of both datasets during the training stage. In all our experiments we fine-tuned the models with a learning rate of 10^{-5}, a decay of 5×10^{-4} and momentum of 0.9, using the default anchor boxes provided by the YOLO V2 model.

Effect of using two distinct datasets in training. As our two datasets have distinct characteristics (number of images, number of instances, variety of poses and occlusions, presence of mostly the face or the body, etc.), in this experiment we evaluate the effect of combining images from both databases (at different proportions) when training the model. To combine the images of both datasets, our method randomly selects an informed proportional amount of images from each dataset during each training epoch. In particular, our goal was to evaluate if the use of "easier" images (in the CelebA dataset) could improve gender detection "in the wild" for harder images present in the Pascal VOC dataset. Although a K-folds validation scheme would be more adequate for evaluating the randomness in the training/test sets and data augmentation, we trained each model only once (for each combination of datasets) due to the relatively high training times of deep architectures.

Table II presents the mean Average Precision (mAP) of our model applied to each dataset separately trained with different proportions. As can be observed, the inclusion of CelebA images (in the proportion of 50%) boosted the mAP of Pascal VOC from 0.69 to 0.74. Although these values might seem low (compared to other detection/recognition problems), we consider them reasonable based on the difficulty of the task "in the wild", as illustrated in Figure 3. It can also be observed that the mAP results for the CelebA dataset did not change much using the additional data from Pascal VOC. As expected, the model trained only using images from CelebA performs poorly in the Pascal VOC dataset, since the latter is a much

more challenging dataset. We consider that training the model with a 50%-50% proportion presents the best overall values, and we use it as our default approach in the next experiments.

Figure 3 presents examples of successful detections (top row) and error cases (bottom row) using the proposed method. A visual analysis of the results shows that our network makes mistakes in bounding boxes that have features of both genders, which can be expected as the anchor box feature map of our network is small and two people can be on the same (x, y) location of the feature map.

TABLE II
RESULTS IN MAP OF OUR 31 LAYER NETWORK USING DIFFERENT PROPORTIONS OF EACH DATASET DURING TRAINING. ALL THE RESULTS ARE AFTER 10,000 ITERATIONS OF FINE-TUNING.

Training Data		Testing	Results		
VOC	CelebA	Data	Man	Woman	mAP
0%	100%	VOC	0.4809	0.4741	0.4775
		CelebA	0.9917	0.9927	0.9922
5%	95%	VOC	0.7114	0.6810	0.6962
		CelebA	0.9922	0.9940	**0.9931**
10%	90%	VOC	0.7490	0.6981	0.7235
		CelebA	0.9914	0.9945	0.9929
25%	75%	VOC	0.7622	0.7222	0.7422
		CelebA	0.9903	0.9939	0.9921
50%	50%	VOC	0.7600	0.7296	**0.7448**
		CelebA	0.9780	0.9881	0.9830
75%	25%	VOC	0.7372	0.7178	0.7275
		CelebA	0.9695	0.9807	0.9751
100%	0%	VOC	0.7136	0.6685	0.6911
		CelebA	0.9044	0.9370	0.9207

Comparison with other gender recognition methods. As other methods deal only with the recognition task, assuming that the person is already localized, they only compute the recognition accuracy (%) of the method using the hits/errors instead of the mAP. To compare our method using this metric, we first select only detections (man or woman) with a prediction confidence larger than a threshold t (set to 0.5 experimentally), and consider it a successful person detection if the IoU w.r.t. the annotated person is larger than 0.5. For those remaining bounding boxes, we compare the ground truth with the detected label, and obtain the final accuracy (in %).

Table III presents the accuracy of our models in the CelebA dataset trained with different strategies. It also shows the accuracy of two state-of-the-art gender recognition methods [4], [12] for the same dataset. As can be observed, our default training protocol leads to an accuracy comparable to state-of-the-art methods for CelebA, and results get slightly better if the model is trained with a combination of CelebA and Pascal VOC images.

Runtimes. We evaluated the runtime of our model using two different GPUs: an Nvidia Titan Xp and an Nvidia Geforce 1080. Our model is able to process a video file with resolution 1056×704 at ~55 FPS on a Titan Xp GPU, and ~38 FPS on a GeForce 1080. When using a camera, the FPS drops to ~26 FPS in both GPUs (the bottleneck is the frame grabber). Training time was around 350 minutes for each model using the Titan Xp GPU, and $10,000$ iterations were processed.

TABLE III
PERFORMANCE COMPARISON (IN %) OF GENDER RECOGNITION ON CELEBA AND VOC DATASETS. OUR RESULTS ARE EVALUATED USING OUR "CLEANED" VERSION OF CELEBA THAT PRESENTS 87 LESS IMAGES THAN THE ORIGINAL CELEBA DATASET.

Method	CelebA
LNets+ANet [12]	98
HyperFace [4]	97
Ours (train = 5% + 95%)	**97.30**
Ours (train = 50% + 50%)	96.04
Ours (train = Celeb Only)	97.13
Ours (train = VOC Only)	87.39

Visualization and analysis. Figure 4 presents additional detection results (only one detection per image), along with the corresponding activation maps using the approach presented in Section IV. It is interesting to note that facial cues are used when the person pose is mostly frontal and the face is reasonably large. On the other hand, features along the whole body are activated in more challenging images, in which the face is not clearly visible. More interesting, the activation maps show that in some cases contextual information is used to infer the gender. More precisely, the motorcycle "votes" for the `man` gender (second and third rows, right column), since there were several images in the training set that contain male subjects close to motorcycles. This is an indication that additional care must be taken to avoid biasing the network.

VI. CONCLUSIONS

In this paper, we presented a real-time gender detection scheme in the wild, validated mostly on the Pascal VOC dataset. Our experiments indicate that by using transfer learning and adding training samples from an "easier" dataset (CelebA) it is possible to obtain reasonable mAP values for gender detection in the wild. In the controlled CelebA dataset, the accuracy of the proposed method is similar to other gender recognition methods that explore mostly facial features.

As additional contributions, we also provide gender annotations for some pedestrian images in the Pascal VOC dataset, and provide a visual analysis of the activation maps for the chosen baseline CNN (YoloV2). The activation maps indicate that facial cues seem to be strongly used when the face is clear in the image, but body cues are useful for more "in the wild" scenarios. They also show that contextual information (such as the presence of other objects in the scene) might bias the detection results.

As future work, we intend to explore temporal information for gender detection in the wild when video sequences are available. We also intend to revise our gender labels for the Pascal VOC dataset by using more human annotators, or by allowing feedback from the community about wrong or dubious labels. Finally, we plan to deal with the cases where both genders are present in the same bounding box.

ACKNOWLEDGMENT

The authors would like to thank the funding agencies CAPES and CNPq. As well as NVIDIA Corporation for the donation of the Titan Xp Pascal GPU used for this research.

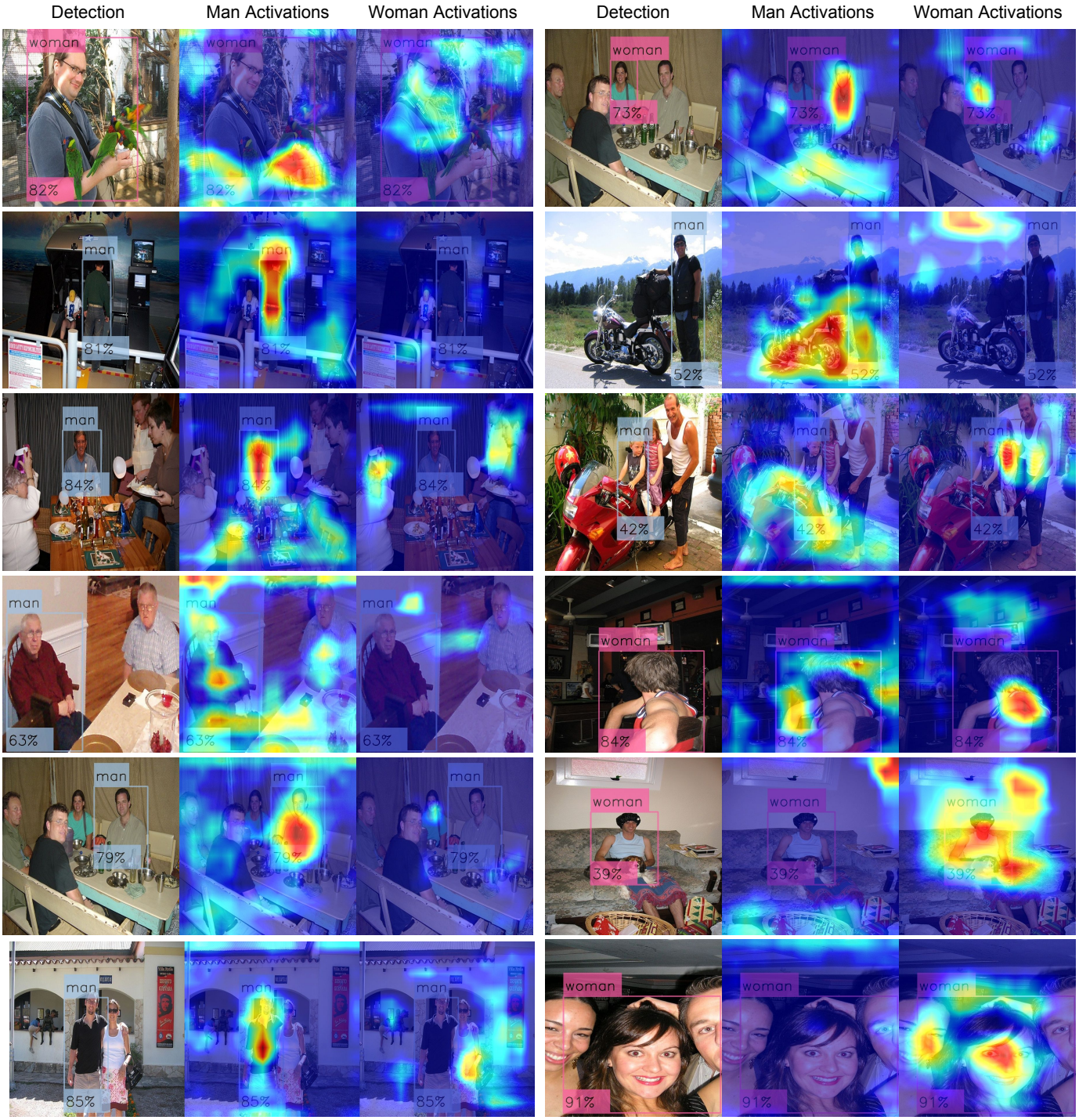

Fig. 4. Examples of visualization of the regions that contributed to the detection in the image.

REFERENCES

[1] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[3] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 34–42.

[4] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *arXiv preprint arXiv:1603.01249*, 2016.

[5] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Technical Report 07-49, University of

Massachusetts, Amherst, Tech. Rep., 2007.

[6] Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian detection aided by deep learning semantic tasks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5079–5087.

[7] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[8] J. van de Wolfshaar, M. F. Karaaba, and M. A. Wiering, "Deep convolutional neural networks and support vector machines for gender recognition," in *Computational Intelligence, 2015 IEEE Symposium Series on*. IEEE, 2015, pp. 188–195.

[9] J. Wang, X. Zhu, S. Gong, and W. Li, "Attribute recognition by joint recurrent learning of context and correlation," *arXiv preprint arXiv:1709.08553*, 2017.

[10] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 618–626.

[11] 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," pp. 98–136, 2015.

[12] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[13] D. Reid, S. Samangooei, C. Chen, M. Nixon, and A. Ross, "Soft biometrics for surveillance: an overview," *Machine learning: theory and applications. Elsevier*, pp. 327–352, 2013.

[14] V. Santarcangelo, G. M. Farinella, and S. Battiato, "Gender recognition: Methods, datasets and results," in *Multimedia & Expo Workshops (ICMEW), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1–6.

[15] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2170–2179, 2014.

[16] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The feret database and evaluation procedure for face-recognition algorithms," *Image and vision computing*, vol. 16, no. 5, pp. 295–306, 1998.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[19] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[21] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2144–2151.

[22] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 789–792.

[23] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[24] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[25] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: visualising image classification models and saliency maps (2014)," *arXiv preprint arXiv:1312.6034*, 2013.

[26] A. Mahendran and A. Vedaldi, "Visualizing deep convolutional neural networks using natural pre-images," *International Journal of Computer Vision*, vol. 120, no. 3, pp. 233–255, 2016.

[27] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*. IEEE, 2016, pp. 2921–2929.

[28] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free?-weakly-supervised learning with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 685–694.

[29] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1713–1721.

[30] R. Rothe, R. Timofte, and L. V. Gool, "Dex: Deepexpectation of apparent age from a single image," in *IEEE International Conference on Computer Vision Workshops (ICCVW)*, December 2015.

[31] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014.

[32] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[33] S. M. Silva and C. R. Jung, "Real-time brazilian license plate detection and recognition using deep convolutional neural networks," in *Graphics, Patterns and Images (SIBGRAPI), 2017 30th SIBGRAPI Conference on*. IEEE, 2017, pp. 55–62.

[34] C. Chen and A. Ross, "A multi-task convolutional neural network for joint iris detection and presentation attack detection," in *2018 IEEE Winter Applications of Computer Vision Workshops (WACVW)*. IEEE, 2018, pp. 44–51.