A convolutional neural network for pedestrian gender recognition

Choon-Boon Ng, Yong-Haur Tay, Bok-Min Goi

Universiti Tunku Abdul Rahman, Kuala Lumpur, Malaysia

{ngcb,tayyh,goibm}@utar.edu.my

Abstract. We propose a discriminatively-trained convolutional neural network for gender classification of pedestrians. Convolutional neural networks are hierarchical, multilayered neural networks which integrate feature extraction and classification in a single framework. Using a relatively straightforward architecture and minimal preprocessing of the images, we achieved 80.4% accuracy on a dataset containing full body images of pedestrians in both front and rear views. The performance is comparable to the state-of-the-art obtained by previous methods without relying on using hand-engineered feature extractors.

Keywords: Gender recognition, convolutional neural network

1 Introduction

Classifying the gender of a person has received increased attention in computer vision research in recent years. There are a number of possible applications, such as in human-computer interaction, surveillance, and demographic collection. While there has been quite a number of works on recognizing gender from facial information alone, less work has been done on using cues from the whole body. In certain situations, using the face may not be possible for privacy reasons, or due to insufficient resolution. Another simpler reason would be that, from the back view of a person, the face is not visible. Most facial gender recognition systems rely on constrained environments, for example frontal or near-frontal view of the head. Thus, we believe there are merits to using the whole body. In particular, this paper focuses on pedestrian gender recognition using computer vision.

Let us consider how we might be able to identify the gender of a pedestrian based on the whole human body rather than relying on the face alone. Due to differences between the male and female anatomy, the body shape can act a strong cue. However, clothing may cause occlusion, such as loose-fitting clothes that make the body shape less obvious. There are clothes for different gender, but similar types are also worn by both, such as long pants or T-shirts. Hairstyle acts a strong cue in the majority of cases, but hair length can be a source of confusion. Despite all this, humans in most situations have the ability to distinguish gender accurately.

The first investigation into gender recognition based on the human body was presented by Cao et al. [1]. Their parts-based method used Histogram of Oriented Gradients (HOG) features to represent small patches of the human body image. These patches are overlapping partitions and used as weak features for a boosting type classifier. Their method gave better classification results that using only raw images with an Adaboost or random forest classifier. Collins et al. [2] proposed descriptors using dense HOG features computed from a custom edge map. This was combined with color features captured from a histogram computed based on the hue and saturation values of the pixels. Guo et al. [3] used biologically-inspired features derived from Gabor filters followed by manifold learning, with linear SVM as classifier. Best results were obtained by first classifying the view (front, back, or mixed) and followed by a gender classifier for each view. Bourdev et al. [4] used random patches called *poselets*, represented with HOG features, color histogram and skin features Their method relied on using a heavily annotated training dataset and context information.

In this paper, we present a discriminatively-trained convolutional neural network (CNN), inspired by the work of LeCun et al.[5] for gender classification of pedestrians in full body images. CNN is a hierarchical neural architecture that integrates feature extraction and classification in a single framework. It is able to automatically learn the features from the training data, instead of relying on the use of hand-crafted features. CNNs have been successfully applied to various pattern recognition tasks such as handwriting recognition [5], face recognition [6], face detection [7], traffic sign classification [8] and action recognition [9].

Our proposed CNN has a straightforward architecture, without incorporating recently proposed architectural elements such as local contrast normalization [10], rectifying nonlinearities [10] and multistage features [11]. Despite its simplicity, we were able to achieve competitive performance comparable to the state-of-the-art for pedestrian gender recognition.

The remainder of this paper is organized as follows. In section 2, the architecture of the convolutional neural network is introduced and explained. In section 3, the dataset used and the details of our proposed CNN is described. The experiment results are presented and compared with other methods in section 4. Finally, in section 5, conclusions are drawn and future work proposed.

2 Convolutional Neural Network (CNN)

Convolutional neural networks are a class of biologically-inspired, multi-layered neural networks. It models the human visual cortex using several layers of non-linearities and pooling. Other classes of such models, inspired by the hierarchical nature of the mammalian primary visual cortex, include the Neocognitron [12] and HMAX model [13]. CNNs are able to automatically learn feature detectors which are adapted to the given task and, to a certain degree, invariant of scale, translation and deformation. The architecture of CNN attempts to realize these invariances using three main ideas, namely local receptive fields, shared weights and spatial subsampling [5].

Traditional CNN typically includes two different kinds of layers inspired by the simple and complex cells in the visual cortex. The convolution layer contains feature maps obtained from convolution of filters with the previous layer's feature maps. The subsampling layer is produced from downsampling each of the feature map.

Figure 1 shows the architecture of our proposed convolutional neural network, which is comprised of 7 layers. The first convolution layer C1 consists of a number of feature maps obtained by convolution of the input with a set of filters, and which are then passed through a squashing activation function. Each unit of a feature map shares the same set of weights for the filter. The connection of each unit to the units located in a small neighbourhood in the previous layer implements the idea of local receptive fields to extract features. The shared weights enable features to be detected regardless of their location in an image. Weight sharing also reduces greatly the number of trainable parameters to achieve better generalization ability.

Let $W_{i,j}$ be the filter of size $n \times m$ which connects the i-th feature map from the previous layer I_i to the j-th feature map C_j and b_j the corresponding trainable bias. The feature map is obtained as following:

$$C_j = \sigma(\sum_{i \in S} W_{i,j} \otimes I_i + b_j)$$

where $\otimes$ denotes the convolution operation and S denotes the set of all or selected feature maps from the previous layer. The squashing activation function σ which introduces non-linearities is normally either a sigmoid $\sigma(x) = 1/(1 + e^{-x})$ or hyperbolic tangent function $\sigma(x) = \tanh(x)$. If the size of a feature map is $h \times w$, then convolution with filter of size $n \times m$ will produce an output of size $(h - n + 1) \times (w - m + 1)$, disregarding border effects.

The layer S2 is obtained by downsampling each feature map in layer C1. In contrast to [5] which uses an averaging operation, we use the so-called max pooling operation [13][14], where the feature map is partitioned into non-overlapping $p \times p$ sub-regions and the maximum value is output from each sub-region. Thus each feature map is downsampled by a factor of p, e.g. the size is halved if $p = 3$. The spatial downsampling operation introduces invariance to small translations. Incidentally, it also reduces the computational complexity for the next convolution layer as the feature map's size is reduced.

In a similar manner, the convolution layer C3 is obtained followed by max-pooling layer S4. Note that each feature map in layer C3 can be either connected to all the feature maps from layer S2 or a subset of it. Layer S4 is fully connected to the units of layer F5, which is a layer of neuron units similar to the hidden layer of a neural network. The output layer contains logistic regression units for classification.

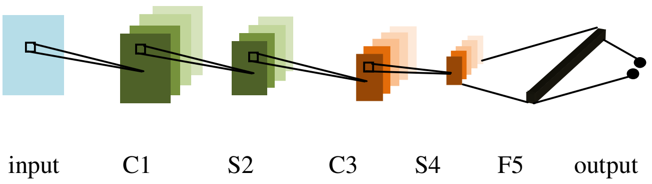


Fig. 1. Architecture of our proposed CNN

3 Experiment

3.1 Dataset

We evaluate the gender classification ability of the CNN on the MIT Pedestrian dataset [15]. There are a total of 924 colour images of male and female pedestrians, in frontal and rear view. The size of the images is 64x128 pixels, with the person's body aligned to the center and the distance between the shoulders and feet approximately 80 pixels. Cao et al. [1] provided gender labels for 888 images (the remaining were indistinguishable) consisting of 600 males and 288 females. The breakdown according to the pose is 420 frontal views and 468 rear views. Figure 2 shows some examples of the images from the dataset.

As preprocessing, the images were cropped to 54x108 by removing the border pixels equally before resizing down to 40x80. Generally better results were obtained compared to without cropping. This could be due to the border pixels containing only background clutter, hence providing miscues. Furthermore, we assume pedestrian detectors would provide a tighter bounding box than compared to the images from this dataset. The images were then converted to grayscale and scaled down to values in the range [0,1] before being used as input to train the CNN.

3.2 The proposed CNN

The detail of the architecture used in our experiments is as follows. Layer C1 contains 10 features maps and uses 5x5 filters, hence when the input image is 40x80, the size of each feature map is 36x76. After downsampling using 2x2 max pooling, each feature map in layer S2 is 18x38. Layer C3 contains 20 features maps with the size 14x34 produced from 5x5 filters. Each feature map in this layer is connected to all the feature maps in layer S2. Layer S4 contains 7x17 feature maps obtained from 2x2 max pooling. All units of the feature maps are connected to each of the 25 neuron units in layer F5. Finally, the output layer has two units for binary classification. We use hyperbolic tangent activations in both the convolution and hidden layers. The total number of free parameters that are learnt by training is 64,857.

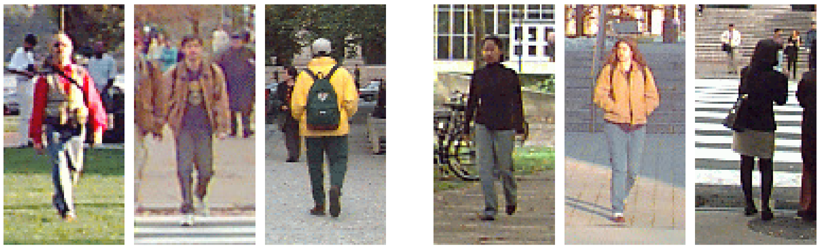


Fig. 2. Examples of pedestrian images from the MIT pedestrian dataset [15].

3.3 Training & Evaluation

Our CNN was implemented using Python with Theano library [16] and trained using mini-batch stochastic gradient descent with learning rate decay. The weights were randomly initialized from a uniform distribution in the range [-√(6/*f*), √(6/*f*)], where *f* equals the total number of input and output connections, following the suggestion in [17].

During each iteration, a batch of images from the training set was presented to the CNN and the weights were updated by backpropagation. Validation was performed using the validation set after each epoch. The minimum validation error was taken as the best result. We used five-fold cross validation and the mean of the validation results was taken to determine the overall measure of accuracy.

4 Results and Analysis

Table 1 shows the results in comparison with other works on gender recognition from human body, evaluated on the same dataset. Our trained CNN achieved an average accuracy of 80.4 % on the validation set which is comparable to the best result by Guo et al. [3]. It should be noted that their method employs a view classifier, followed by a gender classifier for each different view (frontal, rear and mixed). Without using a view classifier, the accuracy is 79.2 %.

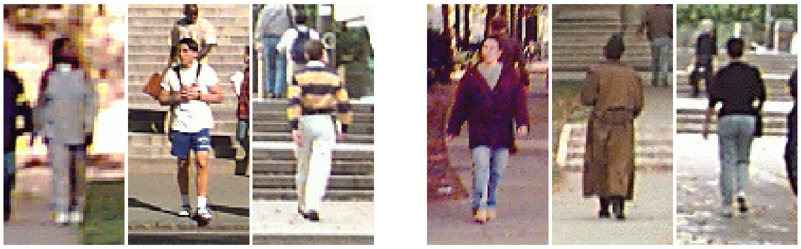
Table 1. Comparison of gender classification accuracy on the MIT pedestrian dataset

Method	Accuracy (%)
Cao et al [1]	75.0
Collins et al. [2]	76.0 (frontal view only)
Guo et al. [3]	80.6
Our method	***80.4***

In contrast, our method goes not require a separate view classifier. The CNN integrates feature extraction and classification in a single framework, where the features are learnt. Furthermore, our CNN uses a relatively small number of feature maps compared to recent works using CNN for recognition tasks [8][18], thus requires less computational intensity.

Interestingly, the method of Guo et al. [3] also used a biologically inspired architecture, with features derived using Gabor filters and max pooling operation. This corresponds to layer C1 and S2 of the CNN. The difference is that Gabor filters can be considered as hard-wired, engineered features, while the CNN learns the optimal features.

Figure 3 shows some examples of classification errors made by the CNN. The first three images on the left are misclassified males and the rest are misclassified females.



(a) M → F (b) F → M

Fig. 3. Examples of misclassified images (a) Male misclassified as female (b) Female misclassified as male

5 Conclusion and Future Work

In this paper, we have presented a convolutional neural network for gender classification of pedestrians in full body images. We achieved 80.4% accuracy on the MIT pedestrian dataset, an accomplishment matching or even better than those using handcrafted features. In our present approach, we use only fully supervised training with a simple, basic CNN architecture. For future work, we plan to implement improvements to our CNN architecture and explore the use of unsupervised pretraining to improve its classification performance.

References

1. L. Cao, M. Dikmen, Y. Fu, and T. Huang, "Gender recognition from body," in *Proceeding of the 16th ACM international conference on Multimedia (2008)*, 2008, pp. 725-728.
2. M. Collins, J. Zhang, and P. Miller, "Full body image feature representations for gender profiling," in *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, 2009, pp. 1235-1242.
3. G. Guo, G. Mu, and Y. Fu, "Gender from body: A biologically-inspired approach with manifold learning," in *Computer Vision--ACCV 2009*, 2010, no. 1, pp. 236-245.
4. L. Bourdev, S. Maji, and J. Malik, "Describing People: A Poselet-Based Approach to Attribute Classification," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2011, pp. 1543-1550.
5. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, p. 2278--2324, 1998.
6. S. Lawrence, C. L. Giles, a C. Tsoi, and a D. Back, "Face recognition: a convolutional neural-network approach.," *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, vol. 8, no. 1, pp. 98-113, Jan. 1997.
7. M. Osadchy, Y. Cun, and M. Miller, "Synergistic face detection and pose estimation with energy-based models," *The Journal of Machine Learning Research*, vol. 8, pp. 1197-1215, 2007.
8. D. Ciresan, U. Meier, J. Masci, and J. Schmidhuber, "A committee of neural networks for traffic sign classification," in *Neural Networks (IJCNN), The 2011 International Joint Conference on*, 2011, vol. 1, no. 1, p. 1918--1921.

9. S. Ji, W. Xu, M. Yang, and K. Yu, “3D Convolutional Neural Networks for Human Action Recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221-231, 2013.
10. K. Jarrett, K. Kavukcuoglu, and Y. LeCun, “What is the best multi-stage architecture for object recognition?,” *Computer Vision, 2009 IEEE 12th International Conference on*, p. 2146--2153, 2009.
11. P. Sermanet and Y. LeCun, “Traffic sign recognition with multi-scale convolutional networks,” in *Neural Networks (IJCNN), The 2011 International Joint Conference on*, 2011, p. 2809--2813.
12. K. Fukushima, “Neocognitron: A hierarchical neural network capable of visual pattern recognition,” *Neural Networks*, vol. 1, no. 2, p. 119--130, 1988.
13. M. Riesenhuber and T. Poggio, “Hierarchical models of object recognition in cortex.,” *Nature neuoscience*, vol. 2, no. 11, pp. 1019-25, Nov. 1999.
14. M. Ranzato, F. J. Huang, Y.-L. Boureau, and Y. LeCun, “Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition,” in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1-8.
15. M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio, “Pedestrian detection using wavelet templates,” in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1997, pp. 193-199.
16. J. Bergstra et al., “Theano: A CPU and GPU Math Compiler in Python,” in *Proceedings of the Python for Scientific Computing Conference (SciPy) 2010*, 2010, no. Scipy.
17. X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS’10)*, 2010, vol. 9, pp. 249-256.
18. Y. LeCun, K. Kavukcuoglu, and C. Farabet, “Convolutional networks and applications in vision,” in *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, 2010, pp. 253-256.