EURECOM
Multimidea Communications Department
Campus SophiaTech
CS 50193
06904 Sophia Antipolis cedex
FRANCE

Research Report RR-15-302

LEARNED VS. HAND-CRAFTED FEATURES FOR PEDESTRIAN GENDER RECOGNITION

April 7th, 2015
Last update April 7th, 2015

Grigory Antipov, Sid-Ahmed Berrani, Natacha Ruchaud and Jean-Luc Dugelay

Tel : (+33) 4 93 00 82 38
Fax : (+33) 4 93 00 82 00
Email : {grigory.antipov,sidahmed.berrani}@orange.com
Email : {natacha.ruchaud,jean-luc.dugelay}@eurecom.fr

[1] EURECOM's research is partially supported by its industrial members: BMW Group Research and Technology, IABG, Monaco Telecom, Orange, Principaut de Monaco, SAP, SFR, ST Microelectronics, Symantec.

LEARNED VS. HAND-CRAFTED FEATURES FOR PEDESTRIAN GENDER RECOGNITION

Grigory Antipov, Sid-Ahmed Berrani, Natacha Ruchaud and Jean-Luc Dugelay

Abstract

This paper addresses the problem of image features selection for pedestrian gender recognition. Hand-crafted features (such as HOG) are compared with learned features which are obtained by training convolutional neural networks. The comparison is performed on the recently created collection of versatile pedestrian datasets which allows us to evaluate the impact of dataset properties on the performance of features. The study shows that hand-crafted and learned features perform equally well on small-sized homogeneous datasets. However, learned features significantly outperform hand-crafted ones in the case of heterogeneous and unfamiliar (unseen) datasets. Our best model which is based on learned features obtains 79% average recognition rate on completely unseen datasets.

Index Terms

Pedestrian gender recognition, convolutional neural networks, HOG, image features, supervised learning

Contents

List of Figures

1 Introduction

An ability to profile people based on their gender is a very important issue which has obvious applications in video surveillance and multimedia retrieval systems. However, quite often it is not possible to get a clear close-shot of a person's face and the gender should be estimated having only a general silhouette of a body. In this work we address the problem of the gender recognition from still images of pedestrians taken in adverse conditions.

The choice of features to describe an object is crucial in computer vision. Existing image features can be roughly divided into 2 categories: the hand-crafted and the learned ones. By hand-crafted features we understand those which are extracted from separate images according to a certain *manually predefined algorithm* based on the expert knowledge. LBP [1], SIFT [2] and HOG [3] features are commonly known examples of hand-crafted features. Contrary to hand-crafted image features, the learned ones are derived from image dataset by a training procedure in order to fulfill a certain task (gender recognition, for example). Convolutional Neural Networks (CNN) [4] are examples of deep neural networks which can be used to extract learned features.

We consider 4 types of features in this study: 2 examples per each considered category. We choose person re-identification [5] features and HOG as examples of hand-crafted features. In order to obtain learned features we train 2 different CNN architectures.

The rest of the paper is organized as following: the bibliography overview and motivations for the choice of features to compare are presented in Section 2; the chosen features are introduced in details in Section 3; the collection of datasets which we have used for experiments is described in Section 4; the performed experiments and obtained results are presented in Section 5 and the main conclusions are highlighted in Section 6.

2 Related work

To the best of our knowledge, PETA collection of datasets [6] is the largest open-access collection of pedestrian images with gender annotation. Authors of PETA propose a universal way to recognize a number of attributes in pedestrian images. Gender is one of the attributes they have considered. In their algorithm, authors of PETA employ re-identification features which were originally described in [5] for the purpose of automatic re-identification of humans in CCTV videos. We use PETA collection of datasets in our experiments. Following the authors of PETA we also choose re-identification features for comparison in our work.[1]

In numerous works authors have found Histogram of Oriented Gradients (HOG) features to be the most appropriate for pedestrian gender recognition. In particular,

[1] We use the re-identification features implementation which was kindly provided to us by authors of PETA.

in [7] authors claim obtaining 76% recognition rate on MIT dataset, while in [8] authors claim obtaining 80% recognition rate on VIPeR dataset.[2] This is the reason why we include HOG features in our comparison as the second example of hand-crafted features.

Due to their exceptional success in recent years, deep CNN have become the first-choice solution for supervised learning on huge datasets [9,10] and [11]. However, CNN are relevant not only in cases when a training dataset is composed of more than a million images. There is a number of examples where CNN were successfully applied to moderate-size training datasets. For example, the famous LeNet5 [12] by LeCun et al. was trained to recognize hand-written digits on $60,000$ training examples while Garcia and Delakis trained a very efficient face finder based only on $3,702$ highly variable face areas [13]. Moreover, in [14] authors employ a CNN for gender recognition on a tiny MIT pedestrian dataset having only 900 images. They claim obtaining 80% recognition rate. In this work, we design our own CNN architecture and corresponding learned features contribute in our comparison.

There are more and more examples ([15, 16]) where a CNN trained on a huge general-purpose dataset is successfully fine-tuned for a very specific classification task with little extra data. Some authors even claim that today fine-tuning of a pre-trained CNN is a must-try method in all image classification tasks [16]. In our work we fine-tune the model which was trained by Krizhevsky et al. [11] on the ILSVRC dataset [17] containing about 1.3 million images.

Contrary to previous works, we do not focus on a single dataset, we rather compare considered features on a versatile collection of pedestrian datasets. Moreover, besides evaluating performance of features on separate datasets, we also compare how well they generalize by applying them on unfamiliar (i.e. completely unseen) datasets.

3 Feature representations

Below we detail the features which are compared in this study.

3.1 Person re-identification features

Person re-identification features of an image [5] is a 2784-dimensional vector which contains low-level colour and texture information. The complete vector is composed of six 464-dimensional vectors each of which is extracted from 6 equal sized horizontal strips from the image. Each strip uses 8 colour channels (RGB, HSV and YCbCr) and 21 texture filters (Gabor, Schmid) derived from the luminance channel. We use a bin size of 16 to describe each channel.

[2] Both MIT and VIPeR datasets are included in PETA collection.

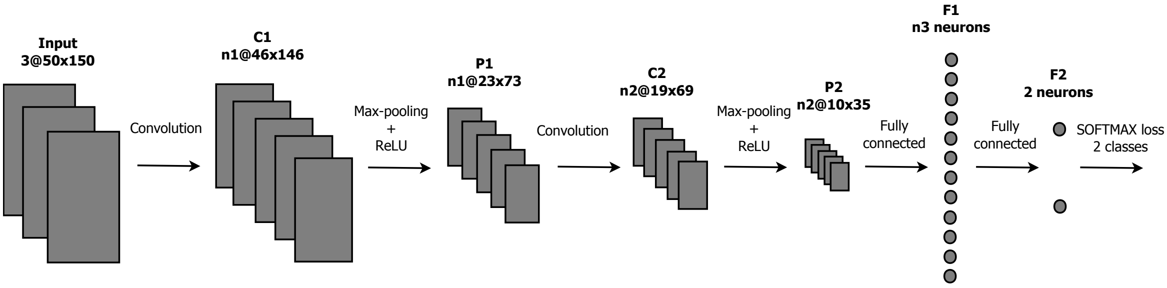

Figure 1: Architecture of *our-CNN*.

3.2 HOG features

In order to extract HOG features, we use 8-by-8 square cells which are organized in 2-by-2 blocks. The number of histogram bins is set to 9. When an image of 50 pixels width and of 150 height is given as input (which is the case in our experiments), the resulting HOG features vector with the described parameters has the dimension 2448.

3.3 Features learned by *our-CNN*

In order to avoid ambiguity, in this section and below, we refer to the CNN which is designed and trained by ourselves as *our-CNN*, whereas the term *AlexNet-CNN* is employed to refer to the CNN which is designed and trained by Alex Krizhevsky.

The architecture of *our-CNN* is presented in Figure 1. *our-CNN* has 2 convolutional layers ($C1$ and $C2$) with 5x5 kernels, each of which is followed by max-pooling with a stride of 2 pixels ($P1$ and $P2$) and Rectified Linear Unit (ReLU) activations.[3] *our-CNN* is ended by 2 fully connected layers ($F1$ and $F2$). The $F2$ layer has 2 neurons (corresponding to the number of classes). The loss is computed by a softmax loss function. As it is depicted in Figure 1, *our-CNN* takes 3 $(50, 150)$-dimensional feature maps (red, green and blue channels of an image) as an input.[4] The first and second convolutional layers $C1$ and $C2$ have $n1$ and $n2$ feature maps respectively. Values of $n1$ and $n2$ depend on the experiment: in the first experiment in Section 5.1 we use $n1 = n2 = 10$, while in the second experiment in Section 5.2 we use a slightly bigger architecture with $n1 = n2 = 20$ (because in the second experiment we have more training images). The number of neurons $n3$ in the last but one fully connected layer $F1$ also varies depending on the experiment: we have $n3 = 25$ and $n3 = 100$ for the first and for the second experiments respectively. In order to augment the size of the training data we use each training image alongside with its mirrored copy. Parameters of *our-CNN*

[3] We have also tried using sigmoid activations but ReLU activations proved to be much faster.

[4] Before being treated by *our-CNN*, all input images are rescaled to $(50, 150)$ size.

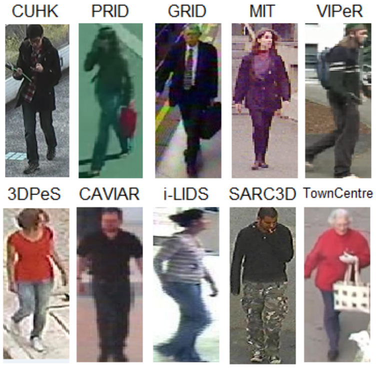

Figure 2: Examples of images from PETA.

have been chosen by trying several architectures and by choosing the most possibly compact one so that performance is not sacrificed.

After training *our-CNN*, we use the obtained weights to calculate the values of neurons in the $F1$ layer for all testing images. These neuron values serve as *our-CNN* features for the testing images. Thus, in the case of *our-CNN* an image is represented by either 25- or 100-dimensional vector (depending on the experiment), which is more than 100 times smaller than sizes of corresponding person re-identification or HOG vectors.

3.4 Features learned by *AlexNet-CNN*

The architecture of *AlexNet-CNN* is described in details in [11]. *AlexNet-CNN* is already a trained model. In our work we only fine-tune it to recognize gender of pedestrians. After fine-tuning of *AlexNet-CNN* we use the weights from the last but one fully connected layer (which contains 4096 neurons) as features for input images. Thus, in the case of *AlexNet-CNN* a features vector has the dimension 4096.[5]

4 Datasets

Originally, the PETA collection consists of 10 datasets of different sizes with a total amount of $19,000$ images. Appearances of images hugely vary between different datasets of PETA in terms of image resolutions (from 17 x 39 to 169 x 365 pixels), camera angles (pictures are taken either by ground-based cameras or by surveillance cameras which are set at a certain height) and environments (indoors or outdoors). Examples of PETA images are presented in Figure 2.

Authors of PETA perform series of experiments on prediction of gender on the whole collection of $19,000$ images [6] using re-identification features. They

[5] Training and fine-tuning of all CNNs in this work is performed using Caffe deep learning framework [18].

Dataset	Train size (♂+ ♀)	Test size (♂+ ♀)
CUHK	$3432 = (2420 + 1012)$	$377 = (189 + 188)$
PRID	$942 = (449 + 493)$	$101 = (50 + 51)$
GRID	$928 = (531 + 397)$	$100 = (50 + 50)$
MIT	$792 = (532 + 260)$	$84 = (42 + 42)$
VIPeR	$1138 = (556 + 582)$	$120 = (60 + 60)$
3DPeS	0	$100 = (50 + 50)$
CAVIAR	0	$68 = (34 + 34)$
i-LIDS	0	$100 = (50 + 50)$
SARC3D	0	$41 = (21 + 20)$
TownCentre	0	$42 = (21 + 21)$

Table 1: Split between training and testing parts per dataset.

randomly split the total collection of images into $9,500$ for training, $1,900$ for validation and $7,600$ for testing. They report obtaining between 79.7% and 81.4% of male gender prediction rate (the results vary depending on the used classifier). We have successfully reproduced this result using several random splits of $19,000$ images in the same proportions as it has been done by authors of PETA. However, the number of unique persons in the PETA collection is much smaller than $19,000$ because PETA contains many images of the same persons which are taken few seconds away from each other by surveillance cameras. Images like that are almost identical and they can considerably bias the resulting prediction rates. After removal of all quasi-identical images from PETA the resulting prediction rates drop down to 63-65$\%$.

This drastic drop in performance proves the importance of the manual filtering of PETA collection. Apart from filtering out images of the same people, we have also removed images with a very low resolution (height is less than 120 pixels or width is less than 40 pixels) and images where a person of interest is not clearly distinguishable (i.e. images of babies in strollers or images of several persons). Finally, we have been left with $8,365$ images (see Table 1) which is less than half as many as the initial size of PETA.

5 Experiments

In order to objectively compare the usefulness of considered features for the gender recognition problem and not to take into account a possible impact of a classifier on resulting prediction rates, we always use the same classifier for all considered features: the SVM classifier with a linear kernel.[6]

In order to evaluate performance of compared models, we use 2 metrics: Mean Average Precision (MAP) and Area Under ROC Curve (AUC) [20].[7]

[6] In particular, we use the publicly available SVM-light implementation of SVM [19].

[7] AUC is an important measure for us because of its invariance to the chosen decision threshold by an SVM-classifier.

Features	**MAP**		**AUC**	
	μ	σ	μ	σ
AlexNet-CNN	0.82	0.05	0.90	0.05
our-CNN	0.79	0.04	0.86	0.03
HOG	0.80	0.04	0.88	0.04
Re-identification	0.59	0.07	0.63	0.09

Table 2: Experiment 1: MAP and AUC statistics calculated over 5 datasets (CUHK, PRID, GRID, MIT and VIPeR).

Features	**Familiar**		**Unfamiliar**	
	MAP	**AUC**	**MAP**	**AUC**
AlexNet-CNN	0.85	0.91	0.79	0.85
our-CNN	0.80	0.88	0.75	0.80
HOG	0.72	0.84	0.56	0.64
Re-identification	0.58	0.60	0.61	0.69

Table 3: Experiment 2: MAPs and AUCs calculated on the set of testing images from "familiar" and "unfamiliar" datasets.

5.1 Experiment 1: evaluation on separate datasets

In the first experiment, we compare the performance of considered features on 5 separate datasets: CUHK, PRID, GRID, MIT and VIPeR. Sizes of training (which is used to train an SVM model) and testing parts for each of these datasets are given in Table 1. In the cases of *our-CNN* and *AlexNet-CNN* the features (i.e. CNN) are learned on the training images as well.

Results of the first experiment are summarized in Table 2. Mean values and standard deviations for 2 considered metrics are calculated over 5 datasets. Based on this experiment, we can conclude that re-identification features are hardly applicable for our problem. Successful results obtained by re-identification features in [6] might be explained by a significant number of quasi-identical images between training and testing datasets (as it is explained in Section 4). The other 3 features show very close performance and none of them can be favoured based only on the first experiment.

5.2 Experiment 2: evaluation of generalization

In the second experiment, we compare the capability of compared features to generalize on heterogeneous and even completely unseen datasets.

Firstly, we train SVM classifiers with different features on the dataset which is composed of training parts of CUHK, PRID, GRID, MIT and VIPeR taken together. Then we test the learned models on the testing parts of the same datasets which are also taken together. Thus, we compare our models on a single big dataset which is composed of non-homogeneous images (different camera angles, environments, etc.)

Results of this experiment are presented in Table 3 (in its "familiar" part). On one hand, we see that for learned features (i.e. *AlexNet-CNN* and *our-CNN*) both MAP and AUC metrics practically coincide with corresponding results in Table 2 (with respect to standard deviations). On the other, hand there is a significant drop of performance for HOG features. These results perfectly make sense because learned features have a possibility to adapt to the heterogeneity of the input data during the training phase, whereas HOG features are not trained and, therefore, less flexible. We do not consider re-identification features in the second experiment due to their poor performance in the first one.

In the second part of the second experiment, we use the same SVM models which have been trained on the collection of training images from CUHK, PRID, GRID, MIT and VIPeR but this time we apply them on the collection of images from completely unseen datasets: 3DPeS, CAVIAR, i-LIDS, SARC3D and TownCentre (see Table 1). In other words, we compare our models on images from "unfamiliar" datasets and, thus, evaluate if they generalize well.

Results of this experiment are presented in Table 3 (in its "unfamiliar" part). *AlexNet-CNN* performs almost equally well as in Table 2 (with respect to standard deviations). *our-CNN* experiences a subtle drop in performance with respect to the first experiment. However, it still shows quite satisfactory recognition rate of 75% despite having only 100 features. The SVM-model learned on HOG features performs poorly on "unfamiliar" datasets showing results which are not so far away from random guessing.

6 Conclusions

In this work we have compared 2 hand-crafted features (HOG and re-identification) and 2 learned features (obtained by *our-CNN* and by *AlexNet-CNN*) in the frame of pedestrian gender recognition problem. Our findings are the following:

1. On small-sized homogeneous datasets HOG features and learned features perform equally well. It complies with previous works ([7, 8, 14]) on pedestrian gender recognition.

2. Learned features significantly outperform HOG features in the case of heterogeneous training data.

3. Contrary to hand-crafted features, both learned features generalize well to completely unseen datasets: MAPs of 79% and 75% respectively.

4. CNN can be used to produce compact feature representations which generalize well (like the one by *our-CNN*).

In our future work, different learned features will be tried in order to recognize other important attributes of soft biometrics (like age, clothing details etc.) We would also like to verify if our findings about learned and hand-crafted features hold in other domains of computer vision. The main requirement is the availability of a versatile collection of datasets like the one that we use in this study.

References

[1] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[2] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE International Conference on Computer Vision*, vol. 2, Toronto, Canada, 1999, pp. 1150–1157.

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, vol. 1, San-Diego, USA, 2005, pp. 886–893.

[4] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, 1995.

[5] R. Layne, T. M. Hospedales, S. Gong *et al.*, "Person re-identification by attributes." in *Proceedings of the British Machine Vision Conference*, vol. 2, Surrey, UK, 2012, p. 3.

[6] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *Proceedings of the ACM International Conference on Multimedia*, Orlando, USA, 2014, pp. 789–792.

[7] L. Cao, M. Dikmen, Y. Fu, and T. S. Huang, "Gender recognition from body," in *Proceedings of the ACM International Conference on Multimedia*, Vancouver, Canada, 2008, pp. 725–728.

[8] M. Collins, J. Zhang, P. Miller, and H. Wang, "Full body image feature representations for gender profiling," in *Proceedings of the International Conference on Compurer Vision*, Kyoto, Japan, 2009, pp. 1235–1242.

[9] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Columbus, USA, 2014, pp. 1701–1708.

[10] I. J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. Shet, "Multi-digit number recognition from street view imagery using deep convolutional neural networks," *CoRR*, vol. abs/1312.6082, 2013.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Proceedings of the Conference on advances in Neural Information Processing Systems*, Lake Tahoe, USA, 2012, pp. 1097–1105.

[12] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[13] C. Garcia and M. Delakis, “Convolutional face finder: A neural architecture for fast and robust face detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 11, pp. 1408–1423, 2004.

[14] C.-B. Ng, Y.-H. Tay, and B.-M. Goi, “A convolutional neural network for pedestrian gender recognition,” in *Advances in Neural Networks–ISNN 2013*. Springer, 2013, pp. 558–564.

[15] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, “Return of the devil in the details: Delving deep into convolutional nets,” *CoRR*, vol. abs/1405.3531, 2014.

[16] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, “Cnn features off-the-shelf: an astounding baseline for recognition,” *CoRR*, vol. abs/1403.6382, 2014.

[17] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *CoRR*, vol. abs/1409.0575, 2014.

[18] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” *CoRR*, vol. abs/1408.5093, 2014.

[19] T. Joachims, “Making large scale svm learning practical,” 1999.

[20] J. A. Hanley and B. J. McNeil, “The meaning and use of the area under a receiver operating characteristic (roc) curve.” *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.