Pattern Recognition Letters 68 (2015) 250–259

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Extracting human attributes using a convolutional neural network approach ☆

Hugo Alberto Perlin [a,*], Heitor Silvério Lopes [b]

[a] Paraná Federal Institute – Paraná, Paranaguá (PR), Brazil
[b] Federal University of Technology – Paraná, Curitiba (PR), Brazil

ARTICLE INFO

Article history:
Available online 26 July 2015

Keywords:
Computer vision
Machine learning
Soft-biometrics
Convolutional Neural Network
Gender recognition
Clothes parsing

ABSTRACT

Extracting high level information from digital images and videos is a hard problem frequently faced by the computer vision and machine learning communities. Modern surveillance systems can monitor people, cars or objects by using computer vision methods. The objective of this work is to propose a method for identifying soft-biometrics, in the form of clothing and gender, from images containing people, as a previous step for further identifying people themselves. We propose a solution to this classification problem using a Convolutional Neural Network, working as an all-in-one feature extractor and classifier. This method allows the development of a high-level end-to-end clothing/gender classifier. Experiments were done comparing the CNN with hand-designed classifiers. Also, two different operating modes of CNN are proposed and compared each other. The results obtained were very promising, showing that is possible to extract soft-biometrics attributes using an end-to-end CNN classifier. The proposed method achieved a good generalization capability, classifying the three different attributes with good accuracy. This suggests the possibility to search images using soft biometrics as search terms.

1. Introduction

In the computer vision community research agenda, extracting high level information from digital images and videos is still an open issue. The main idea is to extract semantically meaningful concepts from images or video, similarly to those that would be extracted and understood by a human. Such procedure, if possible and done automatically (without human intervention) would, certainly, allow a better usage of the huge amount of media (images and videos) currently recorded and stored.

In fact, the interpretation of visual contents can lead to several different outcomes, since it may vary according to the context the observer is immersed. For instance, one might be interested in finding all rectangular or circular shapes present an image, while other might be interested to find complex and highly variable objects, such as cars, animals, or people. An interesting visual content for most people is to identify other people by means of their physical appearance, and this is a key point to various important applications, such as surveillance.

Possibly, the most important and open issue about this extraction process is known as the semantic gap. It represents a kind of distance between the low level information (pixels, edges, shapes, texture) and its high level meaning. Most researchers claim that this information is not available inside the image, but it is observer-dependent. If this is really true, intelligent computational methods are required to deal with that gap.

Biometrics is a research field concerned with the metrics related to human features. There are several types of biometrics ranging, for instance, from physiological (such as fingerprint, DNA, retina) to behavioral (such as voice and gait). In most cases, the acquisition of such kind of biometrics requires the cooperation of the target person [1]. On the other hand, there is another kind of biometric data that can be extracted from images/videos. This kind of information, known as soft biometrics, is related to unspecific human attributes, such as clothing, gender, and ethnicity, for instance. By using these attributes, it is not possible to identify a person unambiguously. However, it is possible to reduce the range of possibilities when searching for a given individual. The great advantage of this kind of biometrics is that the cooperation of the subject is not needed for acquiring the data [2], therefore it fits perfectly for surveillance purposes.

In this paper, the focus is to classify people images according to three different attributes: upper clothes, lower clothes and gender. More specifically, the objective is, giving an image of a person, to identify the type of upper and lower clothes he/she is using, as well as his/her gender. The main motivation for the development of methods to deal with this problem is the improvement of video surveillance systems. Such methods would enable to search a video database

☆ This paper has been recommended for acceptance by G. Sanniti di Baja.

* Corresponding author. Tel.: +55 041 33104697.
E-mail address: hugo.perlin@ifpr.edu.br, haperlin@gmail.com (H.A. Perlin).

http://dx.doi.org/10.1016/j.patrec.2015.07.012
0167-8655/

using high level queries, such as "Find all men with blue t-shirt and black pants", thus saving many hours of human effort to analyze and classify those images.

This is a very hard problem in several ways: there is a high variance in the way that people are dressed as well as in environmental illumination; people can be in many different poses; they also can be partially occluded by other objects or other people; finally, the background in which a target subject is can be complex and different from scene to scene, imposing more difficulties to the identification of the person.

Most classical approaches proposed in the literature are based on a pair: feature extractor and classifier. The former includes a large variety of general/specialized color, texture, and shape descriptors; and the latter uses machine-learning algorithms, such as a Support Vector Machine (SVM). The major problem with these approaches remains in the fact that they are strongly dependent on the human design and, thus, they are far from being done automatically. In this paper a different approach is proposed.

Instead of choosing which kind of feature extractor and classifier will be used, a Convolutional Neural Network (CNN) is employed. This is a deep learning method, where the feature extractor and the classifier are build in a supervised way, tailored according to the nature of the problem. Deep learning methods, such as CNN, have been used as solution to solve some interesting problems in computer vision, especially those related to pattern recognition, such as Optical Character Recognition (OCR), object recognition, pedestrian detection, among others. For some computer vision problems this methodology has achieved the state-of-the-art results.

The main contribution of this work is to create a method to develop an end-to-end supervised classifier capable of extracting high level (semantic) information from images. The content of the paper is as follows. In Section 2, we summarize some results from the recent literature related to the extraction and classification of soft biometrics. In Section 3, a brief review about the concepts of the CNN method is shown. The proposed methodology to extract and classify soft biometrics is presented in details in the Section 4. Next, in Section 5, a hand-designed feature extractor and classifier is reported for comparison purposes. The experiments done and the results obtained are reported in Section 6. The conclusion of the work and future research directions are discussed in Section 7.

2. Related work

Some soft biometrics methods for extracting semantic information from people were developed in the last years. Different methodologies are available taking advantage of certain properties of the problem.

The work of Hansen et al. [3] focused on annotating the humans' features in surveillance videos. A person is described using the primary color of the hair, upper an lower body clothing, as well as his/her height. However, no classification of clothes was done. The proposed methodology includes a background subtraction algorithm, a color descriptor based on the Hue-Saturation-Value (HSV) color model, a height estimator, and a head direction evaluator.

Zhang et al. [4] proposed a methodology for clothes recognition, but restricted only to t-shirts. They present a survey to evaluate which kind of detail/pattern is the most relevant for classifying t-shirts. Based on this survey, some methods were proposed to evaluate the sleeve length, recognize collar and placket, color analysis, pattern recognition and shirt style recognition. More specifically, sleeve length recognition is based on face and skin detector, then, color segmentation and a one-level decision tree classifier.

The development of a robust color detection framework able to identify the colors of clothing in video under real illumination conditions was proposed by D'Angelo and Dugelay [5]. The objective was the identification of hooligans and the prevention of clashes in soccer matches. The proposed algorithm included stages of color constancy, color-space transformation and color matching. The results showed that the proposed methodology was efficient for the specific purpose.

Bourdev et al. [6] proposed a system to describe clothes of people using nine binary attributes. The detection and classification process relies on the Poselet detector, which uses a fully annotated training set. Besides Poselets, a strategy for skin detection and segmentation was also employed. Using a similar method, Weber et al. [7] proposed a clothing segmentation approach. In this case, the H3D dataset was employed to construct a segmentation mask based on the Poselets detection.

Bo and Fowlkes [8] proposed a methodology for pedestrian image segmentation based on hierarchical composition of parts and subparts of image segments. The candidate parts and sub-parts were derived from a superpixels segmentation code. For each candidate segment a score was calculated to determine if it is part of a pedestrian. To do this, authors used a shape descriptor and color and texture histograms. Although the reported results suggested a promising methodology, it turned out to have some drawbacks. The concept of superpixels was also used later by Yamaguchi et al. [9] for clothes labeling, based on pose estimation.

Chen et al. [10] presented a fully automatic system capable of learning 23 binary and 3 multi-class attributes for human clothing. Human pose estimation was performed to find the location of upper torso and arms. Then, 40 features were extracted and quantized. This set of features was used to train a SVM classifier for each desired attribute. A Conditional Random Field (CRF) was calculated to extract the mutuality between attributes.

Employing the well-known Viola–Jones face detector, a modification of the GrabCut segmentation algorithm, the MPEG-7 color descriptor, the HOG shape descriptor and skin color detection, Cushen and Nixon [11] presented a methodology for segmenting the upper body clothes in a mobile platform.

In Dong et al. [12] authors define Parselets as a group of semantic images obtained by low-level over-segmentation, having strong and consistent semantic meaning. With this representation, a deformable mixture parsing model was proposed, based on a set of hand-designed feature descriptors. Using this method those authors managed to parse a human image, obtained by pixel segmentation.

A fully automated clothing suggestion approach was proposed by Kalantidis et al. [13]. The authors used segmentation and hash process to extract and classify the clothes from a digital image. They used color quantization and LBP (Local Binary Patterns) as descriptors, and claimed that the proposed method was scalable and very time-efficient.

Sharma et al. [14] proposed a model for recognizing human attributes and actions. The model was based on a bag-of-words representation and SIFT (Scale-Invariant Feature Transform) features at multiple scales. The results related to human attributes were evaluated as good, but showed that this problem is far from being solved.

Most of the above-cited methodologies are based on the classical approach: some kind of segmentation method and/or hand-designed feature extractors followed by a classifier. In the following sections the main contribution of this paper will be presented: a end-to-end soft biometrics classification framework that needs no segmentation or image pre-processing.

3. Convolutional neural networks

The common process for automatic classification is based, mainly, in two factors, a feature extractor (FE) and a classifier. Usually, the feature extractor consists of hand-designed transformations of the raw input image, seeking to make the classification process more efficient. There are many image descriptors cited in the literature used as features for the classifier. Some usual methods are: HOG (Histogram of Oriented Gradient) [15], Speeded-up Robust Feature (SURF) [16],