Visual Attribute Extraction using Human Pose Estimation

Angelo Nodari, Marco Vanetti, and Ignazio Gallo

Università dell'Insubria, Dipartimento di Scienze Teoriche e Applicate
via Mazzini 5, 21100 Varese, Italy

Abstract. We propose a method to describe how a person is dressed, using an innovative way to extract Visual Information exploiting the Human Pose Estimation. Given the lack of algorithms in this field, we aims to pave the way giving a baseline and publishing a detailed dataset for future comparisons. In particular in this study we show how using the Human Pose Estimation, we are able to extract the essential features for the description of the Visual Attributes. Furthermore, the proposed method is able to manage the problems highlighted in literature regarding the extraction of features from images of people due to their articulated poses. For this reason we also propose a formalization of how describe people's clothing in order to give a starting point and facilitate the analysis and the Visual Attributes extraction phase. Moreover we show how the use of Deformable Structures let us to extract Visual Attributes without the using of segmentation algorithms.

Keywords: Clothing Parsing, Visual Attributes, Human Pose Estimation.

1 Introduction

Over the last years we are witnessing to a growing interest regarding the use of multimedia applications [6] and therefore all the aspects regarding the management of large amounts of image data are becoming of fundamental importance. The methods of indexing and retrieval of multimedia information from these databases are often the critical points in the usage of these applications. For these reasons, research is moving from the study of low-level features to more sophisticated algorithms able to reduce the so called semantic gap with more high level semantic [16].

In particular the information associated with the Visual Attributes of the Object of Interest can facilitate the process of information extraction. For these reasons, the proposed method lies in those algorithms which use low-level features to map high-level information in order to bridge the semantic gap [16].

Our case of study concerns the analysis of the images containing people in particular focusing on their clothing and we propose an innovative procedure to extract Visual Attributes using the Human Pose Estimation.

The information, extracted analyzing how people are dressed, is a very recent field which can be very useful in different areas. In advertising methods, it is possible to match the correct advertisement in according to the information contained in the image [18]. For the surveillance systems, clothing can improve the tracking of people

with the possibility of recognizing them even when the face is not visible or they are turned from the back [23]. There are many works related with the tagging of people in photo collections which take advantages of this kind of information [22] [1], in particular using the information about how a person is dressed permits to manage all the cases where people change their pose and scale, move around and partially occlude each other. Another emerging field is Visual Shopping [14] which is becoming increasingly popular in particular regarding the fashion domain because it is more effective to perform a visual query than type a text to describe a particular cloth.

Regarding the analysis of the segmentation and representation of the clothing in [4] they present a context sensitive grammar in a And-Or graph representation in order to represent different clothes typologies. In particular their aim consist in inferring the composition of a cloth template given an image. In [24] they propose a segmentation algorithm based on a large number of part detectors which is able to separate upper and lower clothing regions. We followed an object-based approach dividing a person into

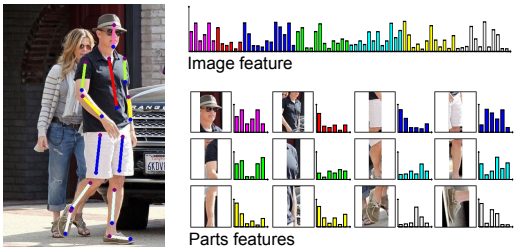

Fig. 1: Starting from an image representing a person (a), we use a Human Pose Estimation approach (b) to extract a set of patches from which we extract the features used to extract the visual information required to build the textual description.

sub regions in order to build a semantic structure understandable by a human. In this way from each of these regions it is possible to extract the pixel-based features which have the characteristic of being related to a particular region and therefore have additional information about their spatial context. Using this information it is possible to easily make reasonings based on a knowledge domain considering the contextual information. The field of object-based image analysis is relatively recent and is mainly used in the field of Geographic information science [12, 13]. Object-based Image Analysis is not intended to replace pixel-based Image Processing, but to enhance it. Its strength consists in the extension of pixel-based methods when they reach their limits and when it becomes important to include semantic concepts and domain knowledge into the analysis process.

We chose to exploit an innovative approach based on the Articulated Human Pose Estimation and we have decided to analyze and use the algorithm proposed in [10, 8]. This method is composed by two parts, a generic detector which uses a weak model of poses and a second step in which the detected parts are pruned by a grabcut model. This method was trained and evaluated starting from a dataset of TV and movie video-shots, but it can be easily adapted to other context, like the one tackled in this study.

After the detection of the area of interest within the image, the next step consists in the extraction of the Visual Attributes. They are firstly formalized and analyzed by [9] who proposed a generative model for learning simple color and texture attributes and [7] which learnt a richer set of attributes including parts, shape, materials, etc. A recent work has highlighted the aspects of Visual Attributes, going beyond the usual features based only on color and texture, trying to extract more semantic information [19]. In [3] they propose a part-based approach to extract information such as gender, hair style and type of clothes using the so called poselets which represents a salient pattern corresponding to a specific viewpoint and local pose and they are used to decompose the aspect of an image to extract the attributes information. This approach obtained good results but is able to extract only a small set of generic visual attributes. In [25] they use for the first time the information related with the human pose to segment different part of the body in order to extract the Visual Information related with the clothing. This information is particularly useful to classify the type of clothes. The segmentation approach is very powerful because is able to give the exact location of a visual attributes and the pixels belonging to, but at the same time is a hardest problem than detect only if a visual attribute is present or not.

In this study, in order to clarify the concept of Visual Attributes, we have classified them into three categories: Object Visual Attributes related with the subparts of the object of interest, which in our case are the different type of clothes; Local Visual Attributes which can be found only in specific locations; Global Visual Attributes which represent more general properties that are typically confined within the boundaries of the object they describe.

For each of these categories, we selected some of the most important attributes and we have evaluated how the proposed method is able to extract them. The proposed method can easily be extended to recognize additional Visual Attributes in this and other domains, by properly selecting the best configuration of all the features to be used.

2 Proposed Method

The proposed method follows a supervised learning approach, and in particular we use a SVM for the classification of visual attributes, consequently the whole process is divided in a training and a test phase. The extraction process of the pattern that describes an image is divided into two main phases: we extract the information on the pose estimation obtaining the so called *Stick Image* Figure 1(a) and subsequently for each sticks of the Stick Image, we extract a feature using Spatial Pyramid Histogram of Oriented Gradient (PHOG) [2], color histograms [17] and raw pixels intensity [21].

The advantage of using a phase of Pose Estimation lies in the possibility to effectively select and manage the position where it is possible to extract the Visual Attributes, managing the problems relating to the different human poses and finally guaranteeing that the extracted features are independent from the body parts rotation, Figure 1(b).

For the Object Visual Attributes Extraction the first step consists in the detection of the subparts which corresponds to the clothes worn by a person and for each of these parts a textual representation is associated, which we use to compose the whole textual description. After the detection of one of this Object Visual Attributes we classify them

in one of the available types using our trained algorithm, for example after have detected the presence of a "Top" Object Visual Attributes we predict its type from the trained classes: camisole, shirt, sweater, etc.

The Global Visual Attributes considered in this work are color and texture [20], which are well suited to characterize particular segments within the image, e.g. the area containing a T-shirt or a skirt. The most important issue related to global attributes is that they require that the object is segmented in order to avoid erroneous inclusion of the background in the computation of the color and texture features, which can significantly decrease the attributes extraction accuracy. Starting from the sticks extracted in the previous step, we want to trace the boundary of the area where we wants to extract the color and texture features. Consequently we have adopted the Pictorial Structures model [11] and considered the human body as a collection of rigid parts that can be connected and therefore in relation with each other. Unfortunately, boundaries of the rigid parts within a Pictorial Structures model does not provide sufficient precision for the purpose of extracting color and texture aims in this study, so we chose to use a more accurate model, called Deformable Structures [26]. They are an extension of Pictorial Structures and are able to capture the non-rigid shape deformations of the human body. Using Deformable Structures, even if the poses of people are complex, the boundaries of the parts to be analyzed are very accurate [26]. Figure 2 shows an example of Pictorial and Deformable Structures applied to an example image, where it is possible to notice how the Deformable Structures provides a much more realistic boundary for the red dress. After the segment identification phase, the information on color and texture is extracted as in the CBIR system proposed in [20]. In this way we avoid to use the common methods for automatic segmentation method which are not enough accurate and reliable for the aim of this study.

Fig. 2: Human body segmentation using Pictorial Structures (a) and Deformable Structures (c). Isolated segments extracted from the images (b) (d).

The local visual attributes are based on two main characteristics: their position is semantically recurrent in the images and for each cloth there is a different set of local visual attributes. Therefore, for each visual attribute we know the position where it is possible to find it, using the Human Pose information. For this reason in the extraction step, for each Visual Attribute, we selected only the areas that contribute to provide a useful information. For example, to test if a pants is short or long, we use the features extracted from the stick of the "left" and "right legs", passing them to the classifier avoiding to use the information about the "head stick" because is not essential in this case. Due to the limited resolution of the images in the dataset, for some types of clothes, some attributes that may be trivial can not be extracted directly from the image. So if we had a very high resolution images, it would be possible to extract more details using the proposed method, which is not a limit of the proposed method

but a limit due to the images resolution. Also for the Local Visual Attributes we have a textual representation that we used in order to build the whole textual description.

3 Experiments

In this section we evaluated the performance of our algorithm in the extraction of the Visual Attributes using Overall Accuracy and K-Value [5].

	Belt	Glasses	Hat	Top Wear	Bottom Wear	Full Wear	Neckline	Sleeves
P.M.	0.23	0.60	0.63	0.53	0.66	0.80	0.43	0.82
S.A.	0.03	0.51	0.42	0.28	0.28	0.34	0.29	0.30
Improvement	+0.20	+0.09	+0.21	+0.25	+0.38	+0.46	+0.14	+0.52

Table 1: Visual Attributes extraction comparison of the proposed method (P.M.) and the state of the art [19] (S.A.), measured using K-value on the Red Carpet Dataset.

We started our experiments using a standard dataset for Human Pose Estimation called Buffy Dataset [10] which consists of a set of images taken from a TV series, with a manual annotation for each image concerning their pose estimation. We have manually annotated each image with a set of label from the Local Visual Attribute classes: Sleeves and Neckline, obtaining the results showed in Figure 3. A basic characteristic of this dataset lies in the fact that people, who are represented in it, are dressed mostly in the same way, but in different poses. Consequently, the results showed in Figure 3, show how the proposed algorithm is able to extract consistent features related to a specific dress according to the change of the people poses with a very high accuracy. We could not test other Visual Attributes given the lack of clothing variety in the dataset.

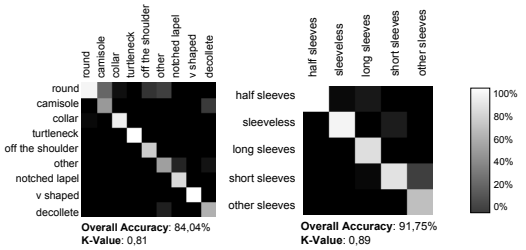

Fig. 3: Confusion Matrix of the Annotated Buffy Dataset for the Neckline (a) and Sleeves (b) classes. The higher is the gray intensity the higher is the accuracy.

Because the main datasets for the analysis of clothes are composed by images extracted by TV series where there is a lack of variability. Moreover other datasets in this particular domain suffers of main problems, as an example the most recent dataset in this field [25] consists only of images of women.

So we built a dataset of images, taken from the web, of celebrities and models, called *Red Carpet Dataset*. This is very challenging because there are men and women, photographed in their everyday life with changes in poses, different brightness, perspective distortions, environment occlusions, etc. This dataset consists of 500 images of different resolutions (from 400px to 1000px), associated with manual annotations related to the

Human Pose and the type of Visual Attributes. We have considered the following Object Visual Attributes: Belt, Hat, Glasses, Top Wear, Bottom Wear, Full Wear; the following Local Visual Attributes: Sleeves, Neckline and the following Global Visual Attributes: Color and Texture. We have evaluated the extraction of these Visual Attributes reported in Table 1. We applied the proposed method on the Red Carpet Dataset and, for each

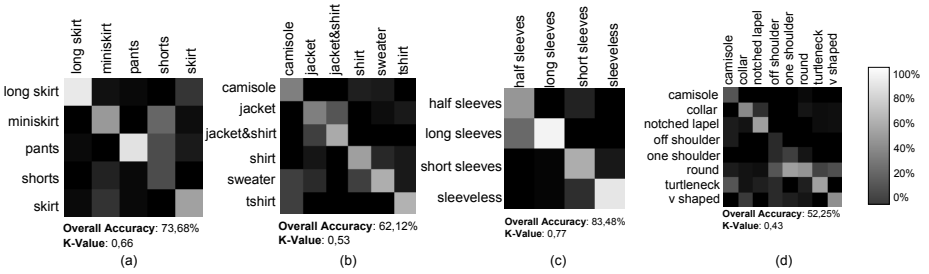

Fig. 4: Confusion Matrix of the Object Visual Attributes Wear Bottom (a), Wear Top (b) and Local Visual Attributes Sleeves (c) and Neckline (d) from the Red Carpet Dataset.

Visual Attribute, we report its confusion matrix. From the results, it is possible to notice that some Visual Attributes are more problematic than others.

We compared the proposed algorithm with the latest algorithm of extraction of Visual Attributes related with the fashion domain [19] on the Red Carpet Dataset showing how this method outperforms the state of the art, the results are reported in Table 1.

Unlike in the dataset [19] in which most of the dresses was not worn and did not appear more than one subject for each image, in this dataset the extraction of the visual attributes is performed on images of people wearing the dress in question and often these people are in complex environments. For these reasons there are still challenges that need to be resolved such as intra-occlusions: arms which overlap the body or are located behind the body and therefore does not remain visible; long hair which often occlude parts of the collar which makes problematic the correct extraction of attributes; etc..

We report the results of the Visual Attributes extraction on the Red Carpet dataset for the Object and Local Visual Attributes in Figure 4 and for the Global Visual Attributes in Figure 5. We report also the computational time evaluated using a single thread C# code, using the OpenCV library [1] on a Intel®Core™i5 CPU at 2.30GHz. They are computed as the average of all the images on the Red Carpet Dataset. Times are different between Visual Attributes because for each of them we selected a different set of Sticks and features. Top Wear 152.27ms, Bottom Wear 230.46ms, Full Wear 500.1ms, Belt 60.15ms, Glasses 32.24ms, Hat 44.68ms, Neckline 271.20ms, Sleeves 186.11ms, Color 37.0ms, Texture 50.0ms.

In this experiment we compared the results in the extraction of color and texture using the Pictorial Structure and Deformable Structure. As can be see from Figure 5, the use of the Deformable Structures allows to increase the performance of extraction for both these Global Visual Attributes.

[1] `http://opencv.willowgarage.com`

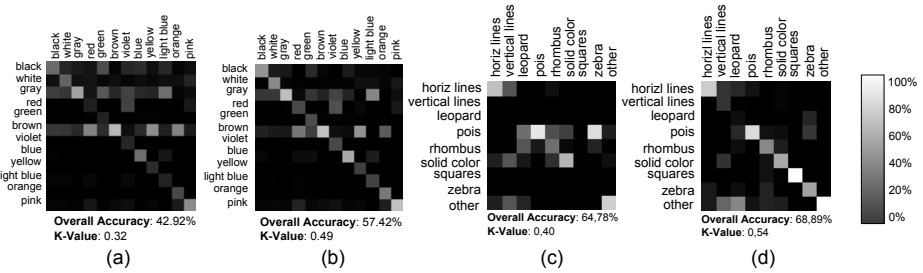

Fig. 5: Global Visual Attribute Colors and Textures confusion matrixes using Pictorial Structures (a,c) and Deformable Structures (b,d) on the Red Carpet Dataset.

The experimental results show how the proposed system is able to describe how a person is dressed with an acceptable accuracy level [15], in order to allow the use of the proposed method in a real application. However, the results also bring to light issues related to intra-occlusions in which there are body parts that are not visible or which occlude key elements for proper recovery of Visual Attributes, such an example the hair which cover the neckline of a shirt, which are the reasons that make the extraction of the Visual Attributes very challenging. For these reasons, we open the way for a future analysis in order to improve the proposed model handling these unsolved problems.

4 Conclusion

In this study we have presented a novel technique that exploits a Human Pose Estimation method to extract Visual Attributes to automatically provide a description of human clothing in a textual way. The robustness of the method was proven by testing, in the experimental section, the individual steps that contribute to the extraction of visual attributes. The major advantage of obtaining a text description of the object of interest consists in the fact that it can be easily indexed and retrieved by a simple textual search query. These results open a wide prospects in the application of this method: context based advertising, surveillance systems, automatic photo tagging and visual shopping, to name a few examples. We hope that this work could pave the way for further research aimed to improve the quality of the results obtained, both for the human pose estimation, in which we have relied on the state of the art, and for the automatic extraction of visual attributes.

References

1. Dragomir Anguelov, Kuang chih Lee, Salih Burak Gktrk, Baris Sumengen, and Riya Inc. Contextual identity recognition in personal photo albums. In *IEEE Conference on In Computer Vision and Pattern Recognition (CVPR)*, 2007.
2. Anna Bosch, Andrew Zisserman, and Xavier Muoz. Representing shape with a spatial pyramid kernel. In *Conference on Image and Video Retrieval*, 2007.

3. Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Describing people: Poselet-based attribute classification. In *International Conference on Computer Vision (ICCV)*, 2011.
4. Hong Chen, Zijian Xu, Ziqiang Liu, and Song Chun Zhu. Composite Templates for Cloth Modeling and Sketching. In *Computer Vision and Pattern Recognition*, 2006.
5. J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.
6. Ritendra Datta, Jia Li, and James Ze Wang. Content-based image retrieval: approaches and trends of the new age. In *Multimedia Information Retrieval*, pages 253–262, 2005.
7. Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *Proceedings of the IEEE Computer Society Conference on CVPR*, 2009.
8. V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Pose search: Retrieving people using their pose. In *Computer Vision and Pattern Recognition*, pages 1–8, 2009.
9. V. Ferrari and A. Zisserman. Learning visual attributes. In *Advances in Neural Information Processing Systems*, December 2007.
10. Vittorio Ferrari, Manuel J. Marn-jimnez, and Andrew Zisserman. Progressive search space reduction for human pose estimation. In *Computer Vision and Pattern Recognition*, 2008.
11. M. A. Fischler and Robert A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, C-22:67–92, 1973.
12. G. J. Hay and G. Castilla. *Geographic Object-Based Image Analysis (GEOBIA): A new name for a new discipline*. Springer, 2008.
13. B. Naga Jyothi, G. R. Babu, and I. V. Murali Krishna. Object Oriented and Multi-Scale Image Analysis: Strengths, Weaknesses, Opportunities and Threats-A Review. *Journal of Computer Science*, 4:706–712, 2008.
14. Lola Khakimdjanova and Jihye Park. Online visual merchandising practice of apparel e-merchants. *Journal of Retailing and Consumer Services*, 12:307–318, 2005.
15. J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, March 1977.
16. Ying Liu, Dengsheng Zhang, Guojun Lu, and Wei ying Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40:262–282, 2007.
17. Wei-Ying Ma and Hong jiang Zhang. Benchmarking of image features for content-based retrieval. In *Asilomar Conference on Signals, Systems Computers*, volume 1, 1998.
18. Tao Mei, Xian sheng Hua, and Shipeng Li. Contextual in-image advertising. In *ACM Multimedia Conference*, pages 439–448, 2008.
19. Angelo Nodari, Ignazio Gallo, and Marco Vanetti. Automatic visual attributes extraction from web offers images. In *Computational Modeling of Objects Presented in Images: Foundamentals Method and Applications*, 2012.
20. Angelo Nodari, Marco Vanetti, Ignazio Gallo, and Simone Albertini. Color and texture indexing using an object segmentation approach. In *EANN/AIAI, CRL publisher*, 2012.
21. Deva Ramanan and Cristian Sminchisescu. Training deformable models for localization. In *Computer Vision and Pattern Recognition*, volume 1, pages 206–213, 2006.
22. J. Sivic, C. L. Zitnick, and R. Szeliski. Finding people in repeated shots of the same scene. In *Proceedings of the British Machine Vision Conference*, 2006.
23. Daniel A. Vaquero, Rogerio S. Feris, Duan Tran, Lisa Brown, Arun Hampapur, and Matthew Turk. Attribute-based people search in surveillance environments. In *In IEEE Workshop on Applications of Computer Vision*, 2009.
24. Michael Weber and Martin Bauml. Part-based clothing segmentation for person retrieval. In *Advanced Video and Signal Based Surveillance*, 2011.
25. Kota Yamaguchi, M. Hadi Kiapour, Luis E. Ortiz, and Tamara L. Berg. Parsing clothing in fashion photographs. In *CVPR*, pages 3570–3577. IEEE, 2012.
26. S. Zuffi, O. Freifeld, and M.J. Black. From pictorial structures to deformable structures. In *Proceedings of the 2012 IEEE*, CVPR '12. IEEE Computer Society, 2012.