International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Review: Parsing Clothes by Giving Weak Supervision using Pose Estimation in Fashion Photographs

Mayuri P. Waghmode[1], Umesh B. Chavan[2]

Department of Information Technology Walchand College of Engineering, Sangli, Maharashtra, India

Abstract: ***In this paper we like to describe the problem of parsing clothing items from images crawled from fashion related websites. The The problem is challenging due to clothing look, layering, fashion, and body shape variation and pose. Presenting the problem of automatically parsing the fashion photographs with weak supervision from the user-generated colour-category labels such as "grey-shirt" and "Black-skirt". Due to immense diversity of fashion items this problem has become very demanding. To solve this problem, the combination of the super pixel-level category classifier learning module, human pose identification module, the Marcov Random Field based colour and category implication module to create multiple well performing category classifiers, which can be applied directly to parse the garments and other items in the images. All the training images are parsed with colour-category tags and the human poses of the images are estimated during the learning phase. Finally, resulting improved pose estimation, providing parsing results with classifier for further use during test.***

Keywords: Markov Random Fields, Fashion Parsing, Weakly supervised learning, SVM.

1. Introduction

Image Parsing can be considered as decomposition of an image I into its visual patterns or categories. For example figure 1 can be parsed and divided into three parts at a coarse level: Grass (background), road (background), Lady. Then Lady can be further decomposed into lady wearing off white long sleeve top, brown shorts, camel shoes .In brief, we can continue decomposing these parts until we reach a resolution limit.

Fashion photograph parsing is a new and emerging research topic in computer vision. Related applications, such as clothes retrieval [3], fashion attribute mining [2], clothes recommendation [20], clothes modelling [5], and clothes classification [4]. These all refer as a great importance in e-commerce websites. Computer vision algorithms to recognize clothing have a variety of impacts, from better social understanding, to improved human identification, surveillance, computer graphics, or content-based image retrieval.

Online clothing shopping is becoming an increasingly popular shopping model. In many online shopping websites such as Amazon.com, Chictopia.com, and myntra.com, customers can conveniently find their favourite clothing by typing some keywords, such as "Blue-Jeans and black-T-shirt". As colour-category tags are ordinary on fashion websites. Fashion parsing means assigning both colour and category labels to every pixel with or without colour-category labels given an image.

For parsing the fashion images with colour and category tags its required to understand three aspects as followed:
1) This approach is weakly supervised as image level tags instead of pixel level tags are available in training phase.
2) The intra-category variations are extremely large for fashion items.
3) The great variations of human poses make the fashion parsing difficult.

This paper also contributes an effective model to recognize and precisely parse pictures of people into their constituent garments, Initial experiments on how clothing prediction might improve state of the art models for pose estimation and a prototype visual garment retrieval application that can retrieve matches independent of pose.

The clothing labelling problem can be considered in two scenarios. First , a constrained labelling problem where we take the users' noisy and perhaps incomplete tags as the list of possible garment labels for parsing, and second where all garment types are considered in dataset collection as candidate labels which can considered as weak supervision .

2. Related Work

In early HASAN, HOGG [6] did this work on Parsing images of persons and classifying them to categories. But, they only included 4 categories namely face, skin, shirt, jacket, tie. There method included an existing MRF formulation incorporating a prior shape model and colour distributions for the constituent parts. Previously, clothing segmentation has depended on using a graph-cut technique.

Later, Kota Yamaguchi [7] also did clothes parsing in photographs there method included 23 categories and 13 colours , they used pixel level colour and category label assignment during training which was time consuming, costly and required manual labour.

In [8] proposed to tackle the clothing parsing problem using a retrieval-based approach. In which parsing is done in two steps: 1) Compute pixel-level confidence from three methods: global parse, nearest neighbour parse, and transferred parse 2) Apply iterative label smoothing to get a final parse.

In [9] this work proposed is almost similar to that of Kota Yamaguchi but instead of using pixel level tags for colour and category they used image level colour and category tags. In this classification considering 13 colours and 23 categories is done, while weakly supervised colour and category labels are provided as input during training phase. This significantly reduces the time, cost and labour.

3. Literature Survey

3.1 Clothing Recognition

There is a large body of research literature on clothing segmentations, modelling and recognition. Hasan et al. [6] and Wang et al. [11] proposed to use different priors to segment clothing. One representative work for clothing modelling is from Chen et al. [10], which used an And-Or graph representation to produce a large set of composite graphical templates accounting for the wide variability's of cloth configurations. Paper [12] proposed to integrate face detection, tracking and clothing segmentation to recognize clothing in surveillance video. The idea behind it is to use latent SVM (Support Vector Machine) to model the relation between different clothing items for classification.

Clothing items have also been used as implicit sign of identity in surveillance scenarios, to find people in an image collection of an event, to estimate occupation, or for robot manipulation. The proposed approach could be useful in all of these scenarios. Parsing clothing can be considered as pre-step for these some applications to boost their performance

3.2 Fashion Parsing

Fashion parsing is getting a comprehensive view of fashion images and understanding them by assigning attributes and category label to pixels in that image. This can be used as a pre-step for applications like clothes retrieval, clothes recommendation, etc.

The relationships between different apparel are modelled by Support Vector Machine (SVM). Clothing recognition and its classification was first explored by yang et al. [12] which used colour segmentation methods and texture features for clothes classification. Later Chen et al. [5] which described apparels appearance with semantic attributes used Conditional Random Field model for classification results. Bossardetal al. [4] proposed a random forest framework to be outperforming as compared to others then they extended it to transfer learning. Recently Chen et al. [13] contributed And-Or Graph as rigorous matching model.

3.3 Weakly Supervised Image Parsing

Now days, images with labels or tags are well known on internet. For, example Image with "Red bag" is present and having a weak label of Red we can locate the presence of bag in that image thereby, improving the performance. Some other work use sparse coding methods like bi-layer sparse coding to parse testing images without training process. For achieving robust label transfer, most considered spatial information between patches.

Yamaguchi used pixel-level labels for category in training while, this approach uses (weak) image level colour-category labels in training phase. The output of Yamaguchi is category label while this gives colour-category labels which are more complete.

Wang et al. [14] presented a process under weakly supervision setting to learn colour attributes and object classes. In this paper method used is based on Marcov Random Field (MRF) inference while their method is based on multiple instance learning.

Yamaguchi [7] used tagged images as input for a clothing parsing method based on fashion image retrieval which was a weak supervision example.[9] uses noisy or incomplete tags for fashion images crawled from fashion websites as weakly supervised data resulting reduced overhead.

3.4 Dataset Construction

There are number of dataset which one can use like MRSC dataset [16] containing 591 images. CamVid dataset [15] contains 711 images. While LabelMe [19] includes 2668 images in its dataset. Yamaguchi [7] Fashionista dataset contains 685 images which is a small scale. Recent [9] uses images from Chictopia.com and provides a huge dataset with colour tags containing 97,490 images. Their dataset can be downloaded from the following website: https://sites.google.com/site/fashionparsing/home .

Dataset used for [0] is much larger and have good quality and visibility. Half of the fashion dataset is used for training while other half is used for testing phase.

Figure 1: Colour and category tags for image from chictopia.com
off White full sleeves Mark Jacobs **Blouse**
brown high waisted Full Romwe **skirt**
Camel Steve Madden **Heel**

Table 1: 13 Colours used for fashion parsing

	beige	*black*	*blue*	*brown*	*gray*	*green*	*orange*
train	31651 (1304)	19673 (1102)	11490 (520)	10139 (833)	3528 (224)	3989 (178)	2014 (111)
test	32085 (1307)	20157 (1118)	10928 (541)	9564 (830)	3296(202)	3960 (178)	1742 (108)
	pink	*purple*	*red*	*white*	*yellow*	*bk*	
train	6666 (328)	2158 (106)	6249 (310)	13592 (793)	5817 (361)	452003 (1341)	
test	6864 (368)	2236 (100)	5879 (302)	13958 (803)	6096 (401)	452338 (1341)	

Table 2: 23 different Categories used for fashion parsing

	face	*sunglass*	*hat*	*scarf*	*hair*
HEAD	3629 (1337)	292 (220)	782 (190)	965 (116)	10806 (1330)
	3675 (1341)	272 (205)	620 (160)	890 (91)	10732 (1332)
	blazer	*T-shirt*	*blouse*	coat	*sweater*
UPPER	3811 (178)	6172 (460)	8669 (474)	3999 (170)	3030 (133)
	3917 (176)	6068 (457)	8198 (461)	4201 (186)	2795 (115)
	jeans	*legging*	*pants*	*shorts*	*skirt*
LOWER	2860 (113)	3124 (123)	3418 (103)	2451 (226)	10214 (438)
	2388 (87)	2699 (110)	2825 (86)	3021 (276)	10353 (445)
	shoe	*socks*	*stocking*		
FOOT	5184 (1300)	242 (83)	1831 (131)		
	5287 (1301)	284 (87)	2316 (157)		
	skin	*belt*	*bag*	*dress*	*bk*
OTHER	26830 (1324)	1056 (390)	6301 (723)	11300 (342)	452003 (1341)
	26690 (1330)	1174 (432)	6692 (739)	11680 (336)	452338 (1341)

3.5 Colour Category Tags

Colour and category tags can be assigned manually or can be crawled from fashion websites (which appear to be noisy or incomplete). In Yamaguchi [7] they have used two Amazon Mechanical Turk jobs to collect annotations. First gathers ground truth pose annotations for 14 body parts while the other gathers ground truth clothing category tags on over-segmented super pixel in Fashion images. Yamaguchi [8] used freely available, weakly annotated Web images available in social networks or fashion sites focused on fashion. In [9] provided fashion all images with colour category tags. They first over-segmented each image into 400 patches, and then associated colour and category tags to each over-segmented region. For example refer Figure 1. They used 13 colours by referring to colour naming research [17] and 23 apparel categories annotation for each patch.

As shown in above Table 1 and 2, 13 colours and 23 categories of labels are used to assign colour-category labels to patches in images which are considered as image level labelling. For each colour and category the numbers of patches in training and testing subset are shown in the first and second row respectively. The numbers of images containing the colour and category are shown in bracket respectively. Face and skin part is divided for better understanding. For multiple colour garments accurate labelling of colour to clothes is difficult, which can be considered as limitation of the current process.

Hasan et al. [6], divide the skin area into "face" and "skin". In the training subset, the number of faces is smaller than the image number, because faces are blocked by black oversize sunglasses or hat sometimes.

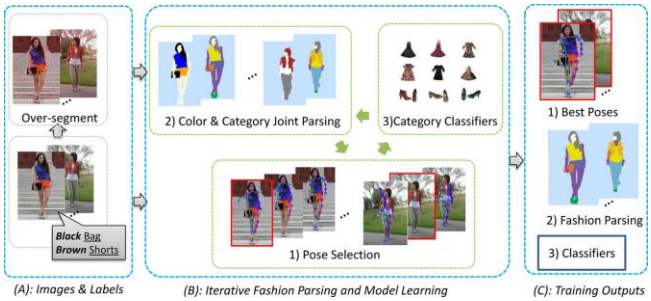

Figure 2: The framework of our weakly-supervised fashion parsing system

Figure 2: (A) Images are over-segmented [18] to decrease computational cost (B) Three components, namely human pose estimation, training images fashion parsing and category classifier training, are iteratively refined. (C) After the iteration converges, the optimal human poses and the fashion parsing results of the whole training image set are obtained. Other outputs of the training process are the well-performing category classifiers which can be used to parse new images.

3.5 Over-segmentation

Over-segmentation of image defines segmenting an image into super-pixels. Yamaguchi [7] used Contour detection with hierarchical image segmentation for generating super-pixels. [18] Proposed a new method for generating super-pixels based on k-means clustering, SLIC, which has been shown to outperform existing super-pixel methods in nearly every respect. In [9] fashion image was segmented into 400 patches and then used for further processing to reduce computational cost as its better to work over super-pixels than pixels.

3.6 Pose Estimation

Both Fashion Parsing and person identification go hand in hand. As fashion garments are hung over the human skeleton. For example, shirt appears at torso region of person while shoes on feet. So, Human pose detection is crucial in Fashion parsing. There are 26 person key points considered such as head neck, right arm, torso, knees, etc.

Pose estimation previously was considered labelling problem like assigning body parts to triangulated region or super-pixels. Current approaches use Conditional Random Field or discriminative model to model human as collection of small parts by decomposing and finding relationship among them. Some also used super pixels, contours, foreground/ background colour model and gradient descriptors to identify human pose.

Yamaguchi [7] used original pose estimation using state of the art flexible mixtures of parts model [16]. Whereas, [9] used state of art full body pose estimator [19]. To further decrease the impact of imperfect pose detection top N-pose detector [20] can also be referred. In Yamaguchi [7] in every iteration pose is recalculated for more perfect results which is time consuming. In [9] top 3 poses are considered and pose with higher confidence score is selected for further processing. This is done only ones in advance for each image. This estimated pose can be further used to incorporate the location feature defined by the current human pose detection.

3.7 Image Patches and Features

In Yamaguchi [8] image features like RGB colour of the pixel, Lab L*a*b* colour of the pixel, MR8 Maximum Response Filters, Image gradients at the pixel, HOG descriptor at the pixel, Boundary Distance and Pose Distance are used. Here, features are imposed over pixel and no patches or over-segmentation is taken into account.

In Yamaguchi [7] feature vector φ consists of normalized histograms of RGB colour, normalized histogram of CIE L*a*b* colour, histogram of Gabor filter responses, normalized 2D coordinates within the image frame,

normalized 2D coordinates with respect to each body joint location.

In [9] after over-segmentation and pose estimation they have acquired location aware features from pose estimation with foreground and background seeding these input is fed to Grabcut algorithm [21]. Once foreground lady is acquired and background is eliminated then over-segmented parches are drawn over the result. Then features like colour, HOG [23], SIFT [22] are taken for each patch and with the location aware feature from pose estimation. Thus, clothes category and colour can be acquired.

4. Fashion Cloth Parsing Model

Hassan [] proposed a shape model containing of a deformable spatial probability for part labelling at each pixel. They made a simple extension to MRF to work simultaneously with multiple objects. Lastly, evaluating the job of segmenting individual categories of apparel in images, depicting people and giving the parsing results.

In Yamaguchi [8] system combines global parse models, nearest-neighbour parse models, and transferred parse predictions. To retrieve similar available apparel as the requested query.

In Yamaguchi [7] parsing clothes is expressed as a labelling problem. In which images are over segmented into super-pixels and then providing category or clothing tags to every segment which is anticipated in CRF model. Then accounting unary potentials for apparel appearance and fashion item location w.r.t body parts. Pair-wise potential includes tag smoothing and fashion item co-occurrence. Pose Estimation is further used to incorporate estimates of clothing items based on location and additional features.

[9] suggest to combine the human pose identification module, MRF based colour-category module and super-pixel level category classifier (learning) module to result multiple category classifiers which can be used to apply on to parse clothing items in images. While in training phase human poses and colour-category tags are estimated in this paper.

5. Conclusions and Future Work

In this paper, parsing of fashion photographs containing colour-category tags are parsed which can be further used for many fashion applications. Given an image and its respective image level colour-category label these frameworks an assign colour-category label to each pixel in that image. Also proposes various classifiers that can be considered to parse a test image with pose estimation to be considered. This also suggests that the weakly supervised fashion tag inputs as reduction in time and cost. This paper also compares various fashion parsing frameworks which are proposed up to date.

In future, structured tags for example, like pattern-category tags (e.g. plaid skirt, stripped T-shirt, etc.) can facilitate fashion clothes parsing. And also consider the challenges of fashion clothes parsing including partial body pose identification, incorporate longer range of fashion garment items, improve performance.

References

[1] S. Liu, J. Feng, Z. Song, T. Zhang, C. Xu, H.Lu, and S. Yan, “Hi, magic closet, tell me what to wear,” in Proc. ACM MM.
[2] T. Berg, A. Berg, and J. Shih, “Automatic attribute discovery and characterization from noisy web data,” in Proc. ECCV.
[3] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan, “Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set,” in Proc. CVPR.
[4] L. Bossard, M. Dantone, C. Leistner, C. Wengert, T. Quack, and L. Van Gool, “Apparel classification with style,” in Proc. ACCV.
[5] H.Chen, A. Gallagher, and B. Girod, “Describing clothing by semantic attributes,” in Proc. ECCV.
[6] B. Hasan and D. Hogg, “Segmentation using deformable spatial priors with application to clothing,” in Proc. BMVC, 2010.
[7] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, “Parsing clothing in fashion photographs,” in Proc. CVPR .
[8] Kota Yamaguchi, M. Hadi Kiapour, Luis E. Ortiz, and Tamara L. Berg, “Retrieving Similar Styles to Parse Clothing,” IEEE Trans.Dd
[9] H. Chen, Z. J. Xu, Z. Q. Liu, and S. C. Zhu. Composite templates for cloth modeling and sketching. In CVPR, 2006. 2, 3
[10] Y. Wang and G. Mori. Max-margin hidden conditional random fields for human action recognition. In CVPR, 2009.
[11] Y. Ming and Y. Kai, “Real-time clothing recognition in surveillance videos,” in Proc. ICIP.
[12] H. Chen, Z. J. Xu, Z. Q. Liu, and S. C. Zhu, “Composite templates for cloth modeling and sketching,” in Proc. CVPR.
[13] G. Wang and D. Forsyth, “Joint learning of visual attributes, object classes and visual saliency,” in Proc. CVPR.
[14] J. B. Gabriel, S. Jamie, F. Julien, and C. Roberto, “Segmentation and recognition using structure from motion point clouds,” in Proc. ECCV.
[15] J. Shotton, J. Winn, C. Rother, and A. Criminisi, “Textonboost: Joint, appearance, shape and context modeling for multi-class object recognition and segmentation,” in Proc. ECCV 2006.
[16] V. D. W. Joost, S. Cordelia, and V. Jakob, “Learning color names from real-world images,” in Proc. CVPR.
[17] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, “SLIC superpixels compared to state-of-the-art superpixel methods,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
[18] Y. Yang and D. Ramanan, “Articulated pose estimation with flexible mixtures-of-parts,” in Proc. CVPR.
[19] D. Park and D. Ramanan, “N-best maximal decoders for part models,” in Proc. ICCV.
[20] R. Carsten, K. Vladimir, and B. Andrew, “Grabcut: Interactive foreground extraction using iterated graph cuts,” inProc.TOG.
[21] D. G. Lowe, “Distinctive image features from scale-invariant key- points,”Int .J. Comput. Vision, vol.60, no.2, pp.91–110, 2004.
[22] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Proc. CVPR.