TRANSFERRING CLOTHING PARSING FROM FASHION DATASET TO SURVEILLANCE

Qi Zheng[1,2], *Jun Chen*[1,2], *Chao Liang*[2,3], *Wenhua Fang*[2,4],, *Xiaoyuan Jing*[1,2], *Ruimin Hu*[1,2,3,4]

[1]State Key Laboratory of Software Engineering, Wuhan University, China
[2]National Engineering Research Center for Multimedia Software, Wuhan University, China
[3]Hubei Provincial Key Laboratory of Multimedia and Network Communication Engineering, China
[4]Collaborative Innovation Center of Geospatial Technology, China

ABSTRACT

In this paper we address the problem of automatic clothing parsing in surveillance video with the information from user-generated tags such as "jeans" and "T-shirt". Although clothing parsing has achieved great success in fashion clothing, it is quite challenging to parse clothing in practical surveillance conditions due to complicated environmental interferences, such as illumination change, scale zooming, viewpoint variation and etc. Our method is developed to capture the clothing information from the fashion field and apply it to surveillance domain by weakly-supervised transfer learning. Most of attribute labels in surveillance images convey strong location information, which can be considered as weak labels to deal with the transfer method. Both quantitative and qualitative experiments conducted on practical surveillance datasets have shown the effectiveness of the proposed method.

Index Terms— clothing parsing, transfer learning, semi-supervised learning

1. INTRODUCTION

Clothing parsing is an image parsing task for labeling garment items on the level of pixels. Though clothing parsing is a relatively new research area in computer vision, it has attracted growing attentions in wide fields, ranging from person identification [1, 2, 3], body shape estimation [4] and content-based image retrieval [5] to fashion images parsing [6]. This paper studies clothes parsing in the surveillance condition, as it can be used as implicit cues of persons identities, locations and even their occupations.

In surveillance domain, related work focused on extracting an entire clothing region [7] or providing labels in image-level [8, 9], which are far from the genuine pixel-level clothing parsing. In contrast, clothing parsing has been widely used in the fashion domain [10, 11, 12, 13] as fashion images are often depicted with kind lighting, standard pose, high resolution and good quality. In this line, a milestone work was done by Yamaguchi et al. [14]. They proposed an excellent framework to parse clothing images into their constituent garments. However, directly using the fashion images as training data to parse surveillance clothing is still challenging because of the intrinsic domain difference between surveillance and fashion conditions. Some failure cases (such as some labels monopolizing the object, missing the human object) in MIT dataset are showed in Fig. 1.

Fig. 1. a. The obviously wrong labeling pictures (the left picture) are more than 15%. b. Pictures parsing with wrong object (the middle and right pictures) account for about 28%.

Some researchers attempt to use transfer leaning to bridge the domain gap between surveillance and fashion domain. The transfer learning aim to store knowledge (called 'trained model' in this paper) gained while solving one problem and applying it (which needs to be updated) to a different but related problem [15]. Recently, Chen et al. [16, 17] used CNN based deep domain adaptation network to model data from two domains jointly, but they only focused on clothing attributes without image segmentation. Shi et al. [18] transferred semantic representation between two domains for person re-identification and search, which only took the coarse segmentation as a latent variable.

Above work inspires us that applying transfer learning on clothing parsing could be a feasible solution to solve the domain gap problem. And exploring an available method to update model is the key to coping with this problem. Then we develop the update method with attribute labels (considered as weak labels here). Unlike previous work, this method can capture the plentiful clothing information from fashion domain. Besides, those image-level labels are easy to achieve.

In our work, we raise an iterative method to parse surveillance images. Firstly, we initialize the model by training the fashion dataset. Then we handle domain adaption problem

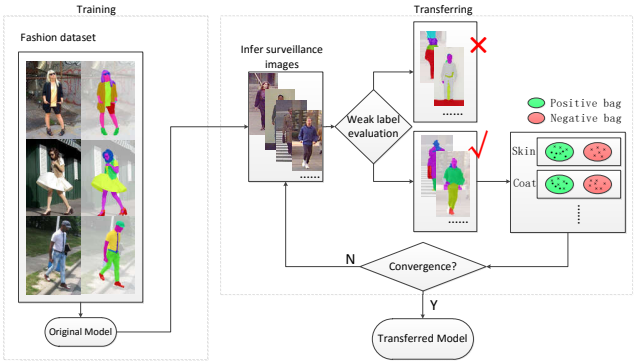

Fig. 2. The left is training the fashion dataset. The right is exploiting pixel-level labels from the elegant results into positive bags (shown by the green cycle) to retrain model.

via multiple-instance transfer learning with weak labels in surveillance dataset as illustrated in Fig. 2. In our experiments, both quantitative and qualitative demonstrate that our method can be applied in a new surveillance dataset.

2. CLOTHING PARSING MODEL

In this section, an improved clothing parsing is described, including general a definition of the clothing parsing, main approaches and the clothing label model.

2.1. Clothing Parsing Definition

Clothing parsing can be considered as a labeling problem, where images are segmented into superpixels and then clothing labels for every segmentation are predicted in a CRF model. In this paper, the probability distribution $P(L|I)$ is given by the model where L, I indicates the set of clothing labels and an image with person respectively. Here we mainly focus on surveillance images with low-resolution, motion blur , which make some labels (like "ring" or "bracelet") unavailable to parse. Thus these labels will be removed from previous fashion framework. Table 1 shows the labels we used.

2.2. Overview of the Clothing Parsing Approach

The clothing parsing is comprised of three steps as Fig. 3 shows:

In the first step, segmentation algorithm [19] is used to obtain the superpixel regions. For a general surveillance image, it typically yields tens of segmentation under the threshold value of 0.05. For the next step, it adapts a implementation [20] to estimate the human pose in the image, which includes a latent variable to infer the typical label of body part. Considering of the pose estimation, the probability distribution $P(L|I)$ is replaced by $P(L|X,I)$, X means $P(L|X,I)$. In

Table 1. Attributes

bag	belt	blazer	hat
blouse	bodysuit	boots	heels
bra	cape	cardigan	jacket
clogs	coat	dress	jeans
flats	gloves	hair	jumper
loafers	panties	pants	leggings
romper	sandals	scarf	shirt
shoes	shorts	skin	skirt
sneakers	socks	stockings	suit
sunglasses	sweater	sweatshirt	t-shirt
tie	tights	top	vest

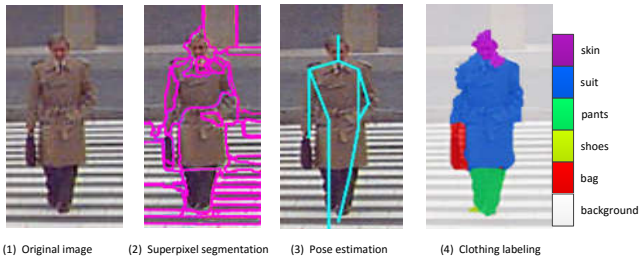

Fig. 3. Overview of clothing parsing

the last step, the $P(L|X,I)$ is modeled with the classic graph model - conditional random field. And the feature vectors exploit the low-level feature combined by normalized histograms of RGB CIE L*a*b* color, Gabor filter responses, normalized 2D coordinates and body joint locations within the image frame.

2.3. Clothing Label Model

In our clothing parsing pipeline, we first initialize our estimating pose $\hat{X}$ and estimating body part type $\hat{T}$ with $P(X,T|I)$ [20] ,the T here is a hidden variable representing a type of pose components. Then the clothing labeling can be estimated by

$$\hat{L} \in argmax_L P(L|\hat{X}, I). \tag{1}$$

The probability distribution $P(L|X,I)$ is modeled with a second order CRF:

$$P(L|X,I) \equiv e^{(\sum_{i\in U}\Phi(l_i|X,I)+\sum_{(i,j)\in V}\lambda_1\Psi_1(l_i,l_j))} \cdot e^{(\sum_{(i,j)\in V}\lambda_2\Psi_2(l_i,l_j|X,I)+\sum_{(i,j)\in M}\lambda_3\Psi_3(l_i,l_j|X,I)-\ln Z)}, \tag{2}$$

where U denotes a region index within a set of superpixels, V is a set of neighboring pairs of image regions, M is a set of mirror part of one's body, λ_1, λ_2 and λ_3 are model parameters, and Z is a partition function. And the unary potential function Φ obtained by

$$\Phi(l_i|X,I) \equiv \ln P(l_i|\phi(s_i,X)), \tag{3}$$

here the s_i is i-th image superpixel region, and ϕ is defined by feature vector as mentioned in section2.2. $\Psi_1(l_i, l_j)$represents a log empirical distribution over pairs of clothing region labels

$$\Psi_1(l_i, l_j) \equiv \ln \widetilde{P}(l_i, l_j). \quad (4)$$

$\Psi_2(l_i, l_j|X, I)$, $\Psi_3(l_i, l_j|X, I)$ estimates the probability of neighboring pairs and the same body type respectively who has the same label:

$$\Psi_2(l_i, l_j|X, I) \equiv \ln P(l_i = l_j|\psi(s_i, s_j, X)), \quad (5)$$

$$\Psi_3(l_i, l_j|X, I) \equiv \ln P(l_i^{'} = l_j^{'}|\psi(s_i, s_j, X)). \quad (6)$$

The graphical model with loopy structure is hard to parse exactly. Therefore, the belief propagation is exploited here to obtain an approximate MAP assignment, using the libDAI implementation.

3. SURVEILLANCE ADAPTATION

In this section, we describe our general technical approach of transferring a fashion based model to the surveillance dataset with some weak labels.

3.1. Overview of the Surveillance Adaptation Approach

Images, in the excellent fashion dataset with pixel-level labels, are taken in ideal fashion conditions. However it is unreliable to directly use the model trained in fashion domain for predicting pixel-level labels in the surveillance domain. In order to bridge this gap, we design a multiple-instance transfer learning for the surveillance adaptation problem as shown in Fig. 2. In our initial clothing parsing task we learn a model based on the fashion dataset. Then we retrain the model with the poisitive/negative bags.

3.2. Surveillance Adaptation Model

A general scheme for a simple optimization can be described as follows. Alternate the following two steps: 1) for a given trained model, find the high confidence images as positive instance. 2) for given positive images, update the clothing parsing model.

The confidence evaluation plays a significant role in the transfer learning model as traditional clustering methods in multiple-instance learning. It is natural to define the function with the position and areas. In our simple scheme, any geometrical information will be utilized to evaluate the performance of clothing parsing without pixel-wise labels. Just as the the hair is always appeared in the upper of the image or the area ratio of shoes can not hold for too much. We define confidence evaluation by

$$c = f(\frac{1}{w_p \cdot D_p + w_a \cdot |a_L - a_0|}) \quad (7)$$

where $f(\cdot)$ is the hyperbolic tangent. The w_p and w_a, computed by discrete degree, denote the weights vector of the samples positions Euclidean distances D_p and areas distance between current area ratio a_L and the mean area ration a_0 respectively. And the results of the clothing parsing will be assigned to positive bags if c is greater than 0.8 (we test the numbers from 0.1 to 0.9 by step 0.1 separately, and 0.8 achieves the best performance).

In the update step, we take the segmentations which are far from the regular position, like edges of an image, as negative samples to ensure negative bags only containing negative ones. Then we train the updated model as section 2 mentioned above.

4. EXPERIMENTAL RESULTS

In the following part, we conduct experiments on two surveillance datasets to investigate the performance of our surveillance adaptation approach.

4.1. Datasets

There are three datasets including **Colourful-Fashion**, **MIT** and **PRID** employed to experiment. Note that the first dataset is used as auxiliary source, and the last two datasets are used as target data. **Colourful-Fashion** [14] includes 685 photos whose pixel-level annotation is provided with 53 different clothing items and 3 additional labels. **MIT** [21] contains 888 pedestrians pictures with 65 attributes [22]including the age, gender or some none-clothing labels. Those attributes are manually reduced to 44 attributes. **PRID** [23] is captured with two different static surveillance camera views. It contains 1134 persons images. To each target dataset, our measurements use 100 images (pixel-labeled by us) for testing and use the remaining for transfer learning in each dataset.

4.2. Clothing Parsing Accuracy

We measure performance of clothing labeling in average pixel accuracy and average IoU (intersection-over-union). Table 2 shows a comparison for 4 versions of our approach in 2 datasets including two surveillance scenes. Obviously the most frequent label present in our images is background. Simply predicting all regions as background results in a reasonably good accuracies (77.6%, 73.3%, 68.6% for each dataset). Thus we use those predictions as our baselines for comparison. The full parsing problem with all 44 garment possibilities is quite challenging. Here the transferred clothing parsing without known items (garment meta-data) is worse than the one with known items, but still performs over the baseline, obtains 78.2% and 72.6% pixel accuracy. The transferred predictions present a efficient improvement than the original ones, achieves an improved 84.2%, 77.7% pixel accuracy for surveillance datasets.

Table 2. Clothing Parsing average Pixel Accuracy Performance. Results are shown over the original model with unknown items (1st line), the original model with known items (2nd line), the transferred model with unknown items (3rd), the transferred model with known items (4th line) and the baseline (bottom).

Dataset	MIT	PRID
Original	0.763	0.708
Original+Items	0.820	0.754
Transferred	0.782	0.726
Transferred+Items	**0.842**	**0.777**
Baseline	0.733	0.686

Table 3. Clothing Parsing average IoU Performance. The IoU shows a relatively low score due to the numerous attributes. But the results with transferred method still outperform the original ones substantially.

Dataset	MIT	PRID
Original	0.082	0.074
Original+Items	0.235	0.189
Transferred	0.113	0.093
Transferred+Items	**0.295**	**0.248**

Table 3 shows the average Iou performance over the 44 garments in the same datasets. Here we consider the original clothing parsing without known items as our baseline, which shows 8.23%, 7.37% average IoU. Note that our method is also roughly obtain 37% ,26% relative improvement with respect to baseline of MIT and PRID datasets. The transferred clothing parsing with known items also perform elegant results with 29.5% and 24.8%.

4.3. Qualitative Evaluation

Our work mainly aims at the surveillance domain. Thus we show some clothing parsing results from both **MIT** and **PRID**. Fig. 4 shows some good parsing results. Our method is able to parse clothing successfully in the challenging resolution, illumination, contrast ratio. Meanwhile, it also can deal with virous orientations, complex background relatively well.

Failure cases are illustrated in Fig. 5. Our approach may lead some wrong results under following scenarios: (a) several persons appear in a single image simultaneously; (b) some items (including clothing garment and background) share the similar appearance; (c) illumination condition is poor.

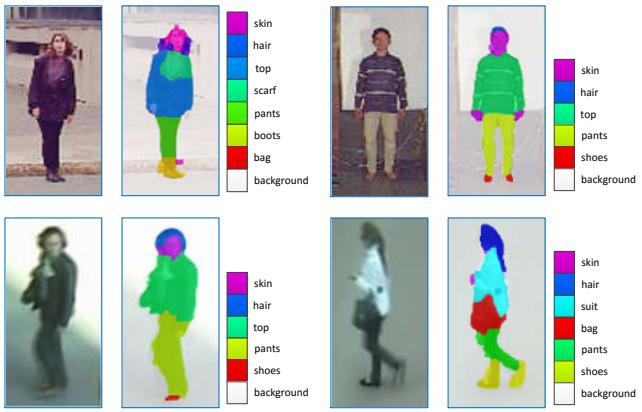

Fig. 4. Some successful parsing results on (1st row) **MIT** and (2nd row) **PRID**.

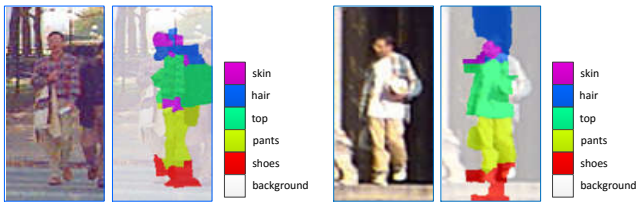

Fig. 5. Some failure cases.

5. CONCLUSION

In this work, we present an iterative parsing optimization method for clothing in surveillance environment. The core idea is: utilizing the multiple-instance to transfer the fashion trained model into surveillance model, with help of weak labels. The method is proved to be effective to parse the clothing without pixel-level labels. The algorithm is simple and the effect promotes significantly. These make the proposed clothing parsing model to be suitable for the surveillance application and the time and manual cost cheaper.

Acknowledgement

The research was supported by National High Technology Research and Development Program of China (2015AA016306), National Nature Science Foundation of China (61231015, 61671336, 61671332, 61562048), Natural Science Fundation of JiangSu Province (BK20160386), the EU FP7 QUICK project under Grant Agreement (PIRSES-GA-2013-612652), the Technology Research Program of Ministry of Public Security (2016JSYJA12), the Fundamental Research Funds for the Central Universities (2042014kf0250, 2014211020203).

6. REFERENCES

[1] Annan Li, Luoqi Liu, Kang Wang, Si Liu, and Shuicheng Yan, "Clothing attributes assisted person reidentification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 5, pp. 869–878, 2015.

[2] Zheng Wang, Ruimin Hu, Chao Liang, Yi Yu, Junjun Jiang, Mang Ye, Jun Chen, and Qingming Leng, "Zero-shot person re-identification via cross-view consistency," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 260–272, 2016.

[3] Mang Ye, Chao Liang, Yi Yu, Zheng Wang, Qingming Leng, Chunxia Xiao, Jun Chen, and Ruimin Hu, "Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2553–2566, 2016.

[4] Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, and Stefanie Wuhrer, "Estimation of human body shape in motion with wide clothing," in *European Conference on Computer Vision*, 2016.

[5] Matthias Weber, Martin Bauml, and Rainer Stiefelhagen, "Part-based clothing segmentation for person retrieval," in *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on*. IEEE, 2011, pp. 361–366.

[6] Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun, "A high performance crf model for clothes parsing," in *Asian conference on computer vision*. Springer, 2014, pp. 64–81.

[7] Andrew C Gallagher and Tsuhan Chen, "Clothing cosegmentation for recognizing people," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[8] Ryan Layne, Timothy M Hospedales, Shaogang Gong, and Q Mary, "Person re-identification by attributes.," in *BMVC*, 2012, vol. 2, p. 8.

[9] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof, "Large scale metric learning from equivalence constraints," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2288–2295.

[10] Agnés Borras, Francesc Tous, Josep Lladós, and Maria Vanrell, "High-level clothes description based on colour-texture and structural features," in *Pattern Recognition and Image Analysis*, pp. 108–116. Springer, 2003.

[11] Hong Chen, Zi Jian Xu, Zi Qiang Liu, and Song Chun Zhu, "Composite templates for cloth modeling and sketching," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. IEEE, 2006, vol. 1, pp. 943–950.

[12] Lubomir Bourdev, Subhransu Maji, and Jitendra Malik, "Describing people: A poselet-based approach to attribute classification," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1543–1550.

[13] Peng Guan, Oren Freifeld, and Michael J Black, "A 2d human body model dressed in eigen clothing," in *Computer Vision–ECCV 2010*, pp. 285–298. Springer, 2010.

[14] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg, "Parsing clothing in fashion photographs," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3570–3577.

[15] Jeremy West, D Ventury, and Sean Warnick, "Spring research presentation: A theoretical foundation for inductive transfer," *Brigham Young University, College of Physical and Mathematical Sciences*, 2007.

[16] Qiang Chen, Junshi Huang, Rogerio Feris, Lisa M Brown, Jian Dong, and Shuicheng Yan, "Deep domain adaptation for describing people based on fine-grained clothing attributes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5315–5324.

[17] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang, "Learning from massive noisy labeled data for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2691–2699.

[18] Zhiyuan Shi, Timothy M Hospedales, and Tao Xiang, "Transferring a semantic representation for person re-identification and search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4184–4193.

[19] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik, "Contour detection and hierarchical image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 5, pp. 898–916, 2011.

[20] Yi Yang and Deva Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1385–1392.

[21] Michael Oren, Constantine Papageorgiou, Pawan Sinha, Edgar Osuna, and Tomaso Poggio, "Pedestrian detection using wavelet templates," in *in Computer Vision and Pattern Recognition*, 1997, p. 193.

[22] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang, "Pedestrian attribute recognition at far distance," in *Proceedings of the ACM International Conference on Multimedia*, 2014, pp. 789–792.

[23] Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof, *Person Re-identification by Descriptive and Discriminative Classification*, Springer Berlin Heidelberg, 2011.