Clothing Extraction using Region-based Segmentation and Pixel-level Refinement

Zhao-Rui Liu
Department of Computer Science and Engineering
Southwest Jiaotong University
Chengdu, China
liuzhaorui1@163.com

Xiao Wu
Department of Computer Science and Engineering
Southwest Jiaotong University
Chengdu, China
wuxiaohk@home.swjtu.edu.cn

Bo Zhao
Department of Computer Science and Engineering
Southwest Jiaotong University
Chengdu, China
dataminer@163.com

Qiang Peng
Department of Computer Science and Engineering
Southwest Jiaotong University
Chengdu, China
qpeng@home.swjtu.edu.cn

***Abstract*—In this paper, we demonstrate an effective method for automatic extracting clothing object from fashion photographs, an extremely challenging problem due to the non-uniform natural backgrounds, various types of apparel and different poses of human models. This method consists of three phases: (1) coarse clothing area localization by pose estimation and superpixel segmentation, (2) region-level image segmentation, (3) pixel-level refinement using spatial information and Grabcut. Experiments on a dataset with 1000 images crawled from Taobao demonstrate that the proposed method outperforms other methods, which can extract clothing from images with complex background.**

***Keywords*-clothing extraction; region-based segmentation; pixel-level refinement; superpixel.**

I. Introduction

With the popularity of e-commerce websites, such as Amazon, eBay, and Alibaba, online shopping becomes a convenient and attractive shopping way for billions of web users. Among them, clothing shopping takes up a large portion of online shopping. Usually, web users, especially female customers, painstakingly spend a few hours each day to browse, search and select clothes that fit their needs. Therefore, an effective application for clothing search by visual similarity would have exceptional value. However, the clothes on e-commerce websites are usually dressed by fashion models to attract customers, which present various colors and styles. In addition, these pictures are taken with natural outdoor background. These properties make clothing visual search a challenging task.

Image segmentation and object extraction has been a hot topic in computer vision and multimedia area. Its target is to identify and extract objects by removing the background and unrelated information. In this paper, the object refers to the clothes and we focus on the clothing extraction from fashion photos. Existing image segmentation methods can hardly extract complete clothing when facing various lighting conditions, diverse colors and styles, and variations in body posture and gesture. In this paper, we propose an effective method for clothing extraction by combining region-based image segmentation and pixel-level refinement. Pose estimation and superpixel segmentation are first used to guide the coarse clothing area localization. Based on the region-granularity clothing segmentation, the rough clothing region is extracted. With the help of clothing integrity and body structure information, the pixel-level segmentation algorithm is performed to achieve better segmentation performance. Experiments on a clothing image dataset crawled from e-commerce website Taobao demonstrates that the proposed approach outperforms the state-of-the-art techniques.

The rest of paper is organized as follows. Related work is first introduced in Section II. The proposed clothing extraction algorithm is elaborated in Section III. The experiments are presented in Section IV and a conclusion is summarized in Section V.

II. Related Work

A. Image Segmentation

Image segmentation is an important step for several image related tasks, such as content-based image retrieval, image annotation and object recognition. There are numerous image segmentation approaches proposed in the past decades, which include threshold based [1], region based [2] and watershed based [3] approaches. In recent years, many graph-based image segmentation approaches are widely used, including normalized cuts [4], efficient graph cut [5] and GrabCut [6]. These algorithms all need to establish an objective function where a minimal value corresponds to the optimal segmentation. Normalized cut [4] computes the eigenvectors of an image which are computationally expensive, so that they are too slow to be suitable for practical applications. The efficient graph-based image segmentation [5] has become one of the most popular approaches due to its good performance in terms of efficiency and effectiveness. As a semi-supervised segmentation approach, GrabCut [6] provides an interactive way. With the assistance of a user-specified bounding box around the object, it estimates the color distribution of the target object and the background using a *Gaussian mixture model*, which achieves impressive performance.

978-1-4799-4311-1/14 $31.00 © 2014 IEEE
DOI 10.1109/ISM.2014.74

B. Clothing Recognition

The vast majority of clothing recognition is based on human detection and pose estimation. In [7], multiple images are used to jointly solve the clothing recognition and segmentation task. These images all contain the identical person wearing the same clothing but are taken from different viewpoints or backgrounds. Graph cuts based on a clothing model learnt from multiple images are used to co-segment clothing images. In [8], a novel multi-person clothing segmentation method is proposed, which can solve the occlusion problem. An automatic system capable of generating a list of semantic attributes for clothes is proposed in [9]. The clothing style rules are modeled by a *Conditional Random Field* on top of the classification predictions from individual attribute classifiers, from which the mutual dependencies between features are explored. A cross-scenario clothing retrieval system is introduced in [10]. To handle the problem of human pose discrepancy, human parsing technique is first used to align human parts. By integration of the online within-scenario similarities with offline calculated cross-scenario similarity transfer matrix, more reliable similarities are obtained, which is used for final cross-scenario image retrieval. Au automatic occasion-oriented clothing recommendation is proposed in [11] to intelligently suggests the most suitable clothing from a given photo album or online shops. The middle-level clothing attributes are adopted to narrow the semantic gap, which are treated as latent variables in SVM based model to provide clothing recommendation. An effective method is proposed to produce an intricate and accurate parse of garment items in [12]. Superpixels and articulated pose estimation are utilized to estimate the clothing classes in a read-world image. In our early exploration, the principal object detection is proposed on top of the efficient graph-based image segmentation, which takes into account both the spatial position and the region size when determining the principal object in a product image [13]. An approach for interactive product image search with complex scenes is proposed [14], which combines the interactive image segmentation for query images, and efficient graph-based principal object extraction for backend image database to extract the foreground objects, respectively.

III. Clothing Extraction

A. Framework

The presence of natural backgrounds and fashion models could significantly influence the performance of clothing image classification and search. In order to identify the clothes in images and reduce the impact of backgrounds and models, we propose an automatic clothing extraction algorithm. The framework is illustrated in Fig. 1. It mainly consists of three phases. For the first phase, a coarse clothing area is first located. As human pose is a useful priori information to help locate the clothes, pose estimation [15] is adopted to parse the head, arms and torso. According to the position of arms and torso, a coarse inner bound and an outer bound are identified, indicating the potential clothing bound and background bound, respectively. Then, the apparel image is segmented into multiple regions using a superpixel segmentation approach. Each superpixel is treated as a region. According to the overlapping size of inner bound and superpixel, regions can be classified into clothing regions and background regions. The clothing regions are considered as the initial clothing area. In the second phase, a region-granularity clothing extraction is undergone to identify the clothes. A region-based segmentation algorithm is proposed, which combines with the color and spatial information to reassign clothing regions and background regions. It can overcome the problem of inaccurate pose estimation. In the third phase, the clothing localization is further refined by taking into account the clothing integrity and body structure information, which improves the performance. In addition, since a superpixel may contain clothing pixels and background pixels at the same time, a pixel-granularity segmentation algorithm, GrabCut [6] is employed to optimize the clothing region at the pixel level. Eventually, the clothing with clear background is extracted from images, which can be used for visual clothing search or classification to improve the performance.

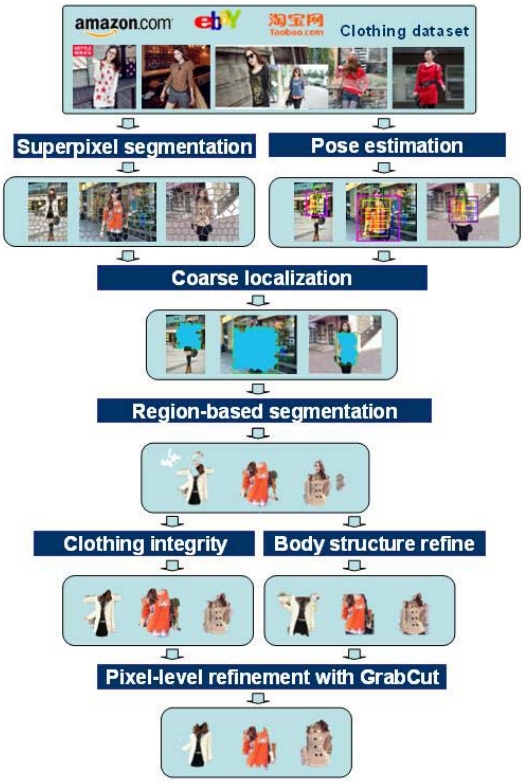


Figure 1. The framework of the proposed clothing extraction approach

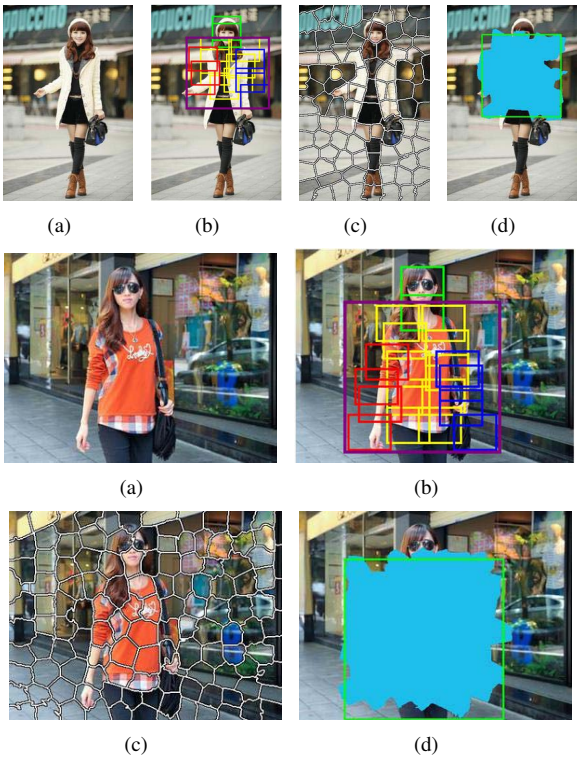


Figure 2. Examples of pose estimation and superpixel results. (a) the original clothing image, (b) detected human parts using pose estimation, (c) detected superpixels, (d) potential clothing area.

B. Coarse Clothing Area Localization

The first step of clothing extraction is coarse clothing area localization. As clothing appeared in e-commerce websites is usually worn by fashion models, the region having clothing is connected with the head and it is located within the regions with arms and torso. Therefore, we exploit the human part parsing technique (i.e., pose estimation [15]) to identify the positions of head, arms and torso. In [15], a mixture of templates is used to capture the orientation of human parts, and a flexible mixture model is utilized to capture co-occurrence relations between parts and standard spring models. This method can reduce the discrepancy caused by human pose variation. Considering that the pose estimation result may be incorrect, a larger rectangular box covering all the regions of arms and torso is treated as the inner bound, which is exploited to to coarsely locate the potential clothing region. Fig. 2(b) depicts pose estimation, with the body parts depicted as colored boxes and the inner bound is illustrated as purple rectangle.

Since superpixel can effectively capture the local redundancy and preserve the boundary of clothing, we combine superpixel and inner bound to outline the coarse clothing area. Meanwhile, superpixels are generated with the method called simple linear iterative clustering (SLIC [16]) by segmenting an image into regions. This method is an adaptation of *k-means* for superpixel generation, which is faster and more memory efficient than existing methods, and exhibits state-of-the-art boundary adherence. A weighted distance is computed by color and spatial proximity, while simultaneously providing control over the size and compactness of the superpixels. By default, the only parameter of this algorithm is *k*, the desired number of approximately equally-sized superpixels. To ensure the speed and effectiveness of the algorithm at the same time, *k* is set as 100 in this paper. The examples of extracted superpixels are illustrated in Fig. 2(c).

Superpixel segmentation is then combined with the inner bound to outline the coarse clothing area. The clothing image I_j is represented as an array $R_j = \{r_1, r_2 \cdots, r_n\}$ with r_i corresponding to ith region, where n is the number of R_j. Then, the initial clothing area/region can be represented as

$$r_i = \begin{cases} clothing & |r_i \cap inner| > \frac{1}{2}|r_i| \\ background & \text{others} \end{cases} \quad (1)$$

C. Region-level Clothing Segmentation

Since the initial clothing area is inaccurate and clothing usually contains a variety of colors and shapes, a robust segmentation algorithm is desired. A region-level segmentation method is proposed. It first assigns regions into five classes. Then, an objective function is then built to represent the similarity among classes. Finally, an iterative method is utilized to minimize the objective function by reassigning regions. The central part is the potential clothing region.

1) Initialize the class label: Intuitively, the region having clothing is usually located in the center of the fashion images, while the background areas with diverse color distributions are spread around the clothing region. In addition, by our observation, regions located in the same corner are relatively similar while different corners usually express significant difference. The situation is more conspicuous when these fashion images are taken on the street with cluttered background. To reduce the effect of noises, the background area is divided into four parts (i.e., top-left, top-right, below-left and below-right) based on the spatial location.

Therefore, given all regions $L_j = \{l_1, l_2 \cdots, l_n\}$ of an image I_j, we will assign each region l_i with one of the labels $\{C_c, C_{tl}, C_{tr}, C_{ll}, C_{lr}\}$, indicating the initial location it belongs to, where C_c is the clothing region, while C_{tl}, C_{tr}, C_{ll} and C_{lr} are the top-left, top-right, lower-left and lower-right corners, respectively.

The Initial label assignment is illustrated in Fig. 3 with different colors. Fig. 3 demonstrates two sets of original images and their corresponding labeled regions. We can see that the assignment results may be incorrect (e.g., Fig. 3(d)) due to wrong pose estimation. To alleviate this situation, we propose a region-level segmentation algorithm to make

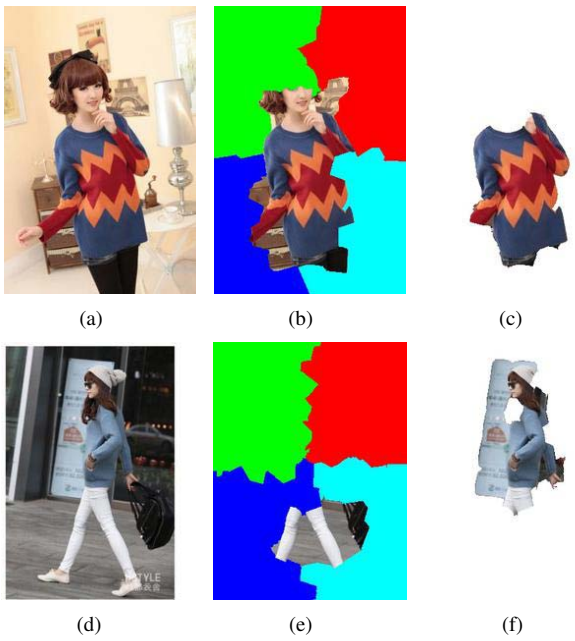


Figure 3. Two sets of images, their corresponding assigned classes and the region-level assignment. Each region will be assigned to either clothing region or background regions (i.e., top-left, top-right, lower-left and lower-right corners), labeled with different colors.

up the drawback of incorrect assignment, which will be elaborated in the following subsections.

2) Construct the objective function: To model the similarity among classes, an objective function is constructed with the sum of weighted squared error $J(\mu, L)$, where μ is a fuzzy factor to control the degree of closeness between regions and classes. Therefore, the segmentation problem is converted to minimize the objective function by reassigning regions to classes and adjusting the fuzzy factor μ. $J(\mu, L)$ is computed as:

$$J(\mu, L) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mu_{i,j}^2 (Dist(r_i, C_j))^2, \ s.t. \sum_{j=1}^{m} \mu_{i,j} = 1 \quad (2)$$

where m is the number of classes. $\mu_{i,j}$ is the fuzzy factor, which represents the degree of closeness between r_i and C_j, and its value is between 0 and 1. $Dist(r_i, C_j)$ depicts the similarity between r_i and C_j. Since C_j is built up by regions with $l_i = j$, we define the similarity between a region and C_j as the weighted sum of the r_i compared to all other regions in C_j. Thus, $Dist(r_i, C_j)$ is computed as

$$Dist(r_i, C_j) = \sum_{r_k \in C_j, k \neq i} e^{-\mu_{k,j}^2} \frac{w(r_i) Dist(r_i, r_k)}{w(C_j)} \quad (3)$$

where $w(r_i)$ and $w(C_j)$ are the number of pixels contained in region r_i and C_j, respectively. $Dist(r_i, r_k)$ represents the similarity between r_i and r_k.

3) Region similarity measure: To measure the region similarity, the most common feature is color. Since describing color with hue, saturation and value is similar to the way of recognizing color by human eyes, the HSV color space is opted as color feature to define clothing color details. We quantize Hue into 8 bins, Saturation into 3 bins and Value into 3 bins, respectively [17], which forms a 72 (8×3×3) dimensional vector *G* with different weights. It is represented as follows [18]:

$$G = 9H + 3S + V \quad (4)$$

In addition to considering the color feature as traditional approaches, we integrate spatial information to compute similarity among regions. Based on our observation, regions closed to Class C_k usually have stronger evidence of belonging to this class than regions far-away. Therefore, when computing the region similarity, we integrate the spatial information to obtain a robust measure. Regions with smaller distance will have higher weights while the regions far away will be assigned smaller weights, which enhances the integrity of clothing. The region similarity $Dist(r_i, r_k)$ is computed as:

$$Dist(r_i, r_k) = exp(S(r_i, r_k)/\sigma^2) C(r_i, r_k) \quad (5)$$

where $S(r_i, r_k)$ is the spatial distance between the center point of r_i and r_k, with pixel coordinates normalized to [0, 1]. $C(r_i, r_j)$ represents the color similarity, which can be achieved by computing the *Euclidean* distance between the color histograms of two regions. σ^2 controls the strength of spatial distance weighting. A larger value of σ^2 will reduce the effect of spatial weight. In our implementation, we use $\sigma^2 = 0.4$.

4) Iteratively segment the clothing: Since the objective function is equality constraints, *Lagrangian* method is used to solve the function. Then, our target is to minimize of the objective function mentioned in Equation (2), with the update of closeness $\mu_{i,j}$ and L. $\mu_{i,j}$ is computed as:

$$\mu_{i,j} = \frac{1}{\sum_{l=1}^{m} [\frac{Dist(r_i, C_j)}{Dist(r_i, C_l)}]^2} \quad (6)$$

The iterative algorithm is composed of the following steps as listed in Algorithm 1. We iteratively update $\mu_{i,j}$ and class labels until classes become stable. As r_i has the same possibility of belonging to five classes, the default value of $\mu_{i,j}$ is set as 0.2.

The result after region-level segmentation is shown as Fig. 3(c) and (f). From this figure, we can see that the clothing can be successfully extracted with the proposed approach. And it can correct and identify the clothing region even though the initial pose estimation gives wrong result. As mentioned above, regions are bounded to each class based on $Dist(r_i, C_j)$, which is computed by μ and L, representing the fuzzy behaviour of this algorithm.

Algorithm 1 Region-based segmentation

Input:

C_j $(j = c, tl, tr, ll, lr)$ and L

Output:

Clothing regions and background regions

Procedure:

Set $m = 0$, $\mu_{i,j}$ = 0.2, initialize $\mu = [\mu_{i,j}]$ matrix
Set input C_j as C_j^m
repeat
  $m = m + 1$
  Update $\mu_{i,j}$ by Equation (6), get μ^m
  Update L by $l_i = \underset{j}{argmax}(\mu_{i,j})$
  Refresh C_j^m based on L
until $C_j^{m-1} = C_j^m$

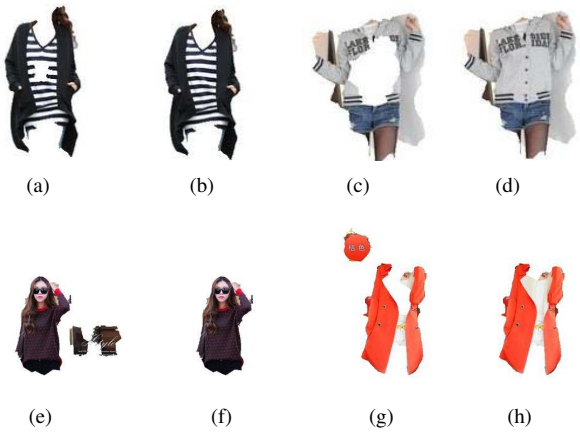
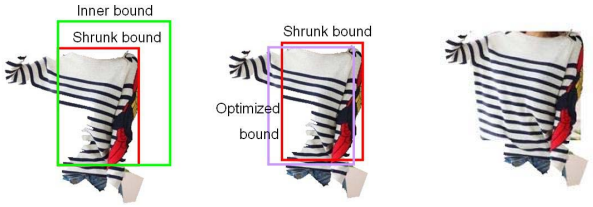


Figure 4. Refined results based on clothing integrity. (a), (c), (e) and (g) are four images with region-based image segmentation. (b) and (d) are results after pixel fill, while (f) and (h) are results after fragment removal.

D. Refine the clothing at pixel-level

1) Refinement by clothing integrity and body structure information: Since the region-based segmentation algorithm ignores the clothing integrity and the inner bound located by pose estimation, which contains human body structure information, we will exploit spatial and inner bound information to optimize the segmentation result. This process mainly consists of three steps: enclose the background pixels which are surrounded by clothing regions, remain the largest region and remove unconnected small parts, and adjust the inner bound for clothing area localization.

The first step is called as pixel fill. Taking into account the integrity of a clothing, the pixels surrounded by clothing regions should belong to the clothing. Thus, we fill the clothing area by assigning the interior pixels to C_c. The second step is called fragment removal. For the region-based method, the foreground area may contain several unconnected parts with one dominant large region and some small fragments. If regions in the background have pretty



(a) Shrunk bound (b) Optimized bound (c) Refined result

Figure 5. Refined result with body structure information. Inner bound is the bounded region with pose estimation result. Shrunk bound is the reduced bound, while optimized bound is the moving bound with the maximum overlap.

similar color distribution with the clothing object, it will cause this situation, which in inappropriate. For this case, We will remain the largest part as clothing area, and small fragments will be filtered simultaneously. The refined results by considering clothing integrity is illustrated in Fig. 4. Holes within the clothing region are filled after pixel fill step, which makes the foreground region become a whole part. While using fragment removal, some unrelated small components are removed, so that noises are eliminated.

As mentioned above, the initial inner bound is determined by arms and torso position, which contains body structure information. By observing Fig. 3, we know that the initial inner bound may locate wrong areas due to false pose detection, which makes the inner bound information becomes meaningless and misleading. We combine the initial inner bound and region-based result to improve the performance. We firstly shrink the inner bound until it connects with the region-based result. The reduced inner bound is identified as potential clothing area. To minimize the error, we move the reduced bound to achieve the maximum overlap with the region-based segmentation result, which is defined as follows,

$$bound = argmax(|bound \cap clothing|) \tag{7}$$

Then, $|bound \cup clothing|$ is the refined clothing result. If there is no overlap between the inner bound and the region-based segmentation result, we do not adjust the inner bound. The refined result and the optimized bound integrated with body structure information is shown in Fig. 5.

2) Pixel-level refinement with GrabCut: Since the size of superpixel is almost fixed, a candidate clothing region also contains many background pixels. Therefore, a pixel-level segmentation is needed to further optimize the image segmentation result. As a highly influential work, GrabCut [6] segments the images by optimizing the energy function iteratively at pixel level, and considers both texture and edge information, which achieves the state-of-the-art performance. Moreover, it supports the user interaction to input a mask around the potential object. We deploy GrabCut to

(a) (b) (c)

Figure 6. The extracted clothing regions with GrabCut for further refinement. (a) original images (b) refined images with spatial information (c) pixel further refined result with GrabCut.

optimize the segmentation result at pixel level, and combine it with our refined result to achieve superior performance.

Instead of initializing GrabCut manually, we automatically initialize the algorithm by refined result mentioned in above subsection. The pixels belonging to candidate clothing region are labeled as unknown part, while other pixels are labeled as background. The refined clothing area is initially used to train clothing color models, which helps to determine the foreground at pixel level. The other pixels except clothing region are initially set as hard background labeling. They are input into GrabCut as a mask to obtain the final clothing extraction result. The extracted clothing is shown in Fig. 6. From this figure, we can see that although the refined results already have good performance, many background pixels are falsely treated as the clothing objects. The extracted results are further refined with GrabCut. Some backgrounds pixels are eliminated from the clothing region.

IV. Experiments

A. Dataset and Performance Metrics

There is no public clothing image dataset and corresponding ground truth available for evaluating the performance of clothing extraction. We crawl 1000 images from Taobao[1] with a great diversity in clothing appearance, backgrounds, poses and lighting conditions, which act as the dataset for performance comparison. All images in the dataset have human models. Five students manually segment each image and extract the clothing part at the pixel level as the ground truth.

To evaluate the performance, *Precision*, *Recall*, *F-Measure* and *Accuracy* are used as the performance metrics to measure the performance of the clothing extraction algorithm at pixel level. The accuracy is measured by the intersection-over-union metric ($\frac{GT_i \cap R_i}{GT_i \cup R_i}$), where GT_i is the ground truth of the clothing region of I_i and R_i represents the extracted clothing. It is a standard metric in PASCAL VOC challenge[2]. *F-Measure*, *Precision* and *Recall* are defined as follows:

$$F = \frac{2 * Precision * Recall}{Precision + Recall} \tag{8}$$

$$Precision = \frac{\#\ of\ correctly\ detected\ clothing\ pixels}{Total\ \#\ of\ detected\ clothing\ pixels} \tag{9}$$

$$Recall = \frac{\#\ of\ correctly\ detected\ clothing\ pixels}{Total\ \#\ of\ clothing\ pixels} \tag{10}$$

Table I
THE RESULT OF CLOTHING SEGMENTATION

	Precision	Recall	F-measure	Accuracy
GrabCut [6]	**85.22**%	65.98%	74.37%	59.20%
POD [13]	59.86%	91.20%	72.28%	56.59%
Scut [19]	53.96%	**92.13**%	68.06%	51.58%
Our method	81.63%	82.71%	**82.16**%	**70.52**%

B. Experimental Results

To evaluate the performance, we compare the proposed method with the *Principal Object Detection (POD)* [13], *Saliency cut (Scut)* [19], and *Interactive Grabcut* [6]. The target of POD is to extract the principal object in clothing images, which is induced from a simplified *GMM* and the efficient graph-based image segmentation. Based on the intuition that the object should be in the middle of the image and the size should not be small, the component in the middle and with large region will be treated as the clothing object. Scut achieves the state-of-art segmentation performance. It firstly uses *Region Contract (RC)* method to get a saliency map. Then, the saliency map is binarized with a fixed threshold. At last, iterative GrabCut is utilized to extract object. In our implementation, we set the binarized threshold as 55, which has achieved 95% recall rate and is proved useful in [19]. GrabCut is an interactive image segmentation solution with human interaction by dragging a rectangle region in the query image to guide the object identification. For better evaluate the performance of automatic clothing extraction, the initial inner bound obtained from pose estimation is treated as the initial rectangle for GrabCut.

Table I lists the performance comparison of the proposed method with baseline approaches. Overall, the proposed approach achieves the best performance in terms of F-measure and accuracy compared to other methods. Although baseline methods work quite well in other datasets without clothing, their performances decline rapidly when facing clothing extraction. Because most fashion images are captured outdoors with cluttered background and different lighting conditions, and the fashion models have inconsistent poses and gestures,

[1]http://www.taobao.com

[2]http://pascallin.ecs.soton.ac.uk/challenges/VOC/

(a) Original (b) POD (c) Scut (d) GrabCut (e) Our method (f) Ground truth

Figure 7. Examples of extracted clothing with different segmentation algorithms. We can see that the proposed method achieves good performance, which is pretty closed to the ground truth.

it makes the clothing extraction a challenging task. It significantly affects the extraction performance of POD, Scut and GrabCut. With the guidance of inner bound, GrabCut [6] has the best precision. Unfortunately, it has pretty lower recall compared to other methods. Due to potential detection error of human body detection and pose estimation, the initial inner bound input to GrabCut may be inaccurate, which may only covers a small amount of clothing region, like the example in Fig. 3(e), or even totally wrong. Given a bound region, the GrabCut has excellent capability to segment the foreground object. Therefore, the GrabCut has a pretty high precision while relatively low recall. Baseline methods POD [13] and Scut [19] have the similar performance, which have high recall values. POD is affected by the performance of efficient graph-based image segmentation and complex backgrounds. Some clothing regions are mixed with the background areas, or background regions are falsely treated as foreground clothing, making them form a large region mixed with foreground and background. These regions have higher weights because of the large region size. Therefore, the recall is high while precision is low. Scut performs poor when clothing is not saliency object in images and clothing contains a variety of colors. Thus, the extracted object by Scut might not be the clothing at all. On the contrary, the proposed method first locates the potential clothing region with body detection and pose estimation, and then integrates region-level image segmentation and pixel-level refinement using spatial information and GrabCut. It can compensates the shortcomings of existing approaches for clothing images. With the three stage segmentation, our method can accurately extract the clothing region from fashion images with cluttered background. Even the initial inner bound detected by pose detection is not accurate, the proposed method can correct it in the following steps. This method achieves a balance between precision and recall. Both of them are higher than 80%, which can be applicable for automatic clothing extraction for backend shopping image datasets with complex backgrounds.

Fig. 7 shows three examples of the extracted clothing with different segmentation approaches. From this figure, we can see that the performance detected with POD is relatively good. Although the major parts of the clothes can be extracted, the result is not perfect. Either nearby backgrounds are falsely included or some parts of clothing object are neglected (see Fig. 7(b)). For Scut, there are large portion of backgrounds are falsely detected as the clothing object (see Fig. 7(c)). Due to the affect of initial inner bound derived from the pose detection, the extracted clothes are usually incomplete using GrabCut (see Fig. 7(d)). Some parts belonging to the clothing object is missing. Generally, the extracted clothes using our method achieves good performance, which is pretty closed to the ground truth (see Fig. 7(e) and (f)). The extracted clothes are more complete and meaningful compared to other approaches. From these examples, we have a couple of interesting observations. Firstly, our algorithm can correctly extract clothes when human models have different pose and gesture. They can be either standing, sitting or even stooping. Secondly, the images with complex backgrounds are correctly extracted, which is a challenging problem for traditional segmentation methods. The presented framework is fast and scalable, allowing our clothing segmentation automatic and effective.

V. Conclusion

In this paper, we explore the clothing extraction algorithm with the combination of region-based clothing segmentation and pixel-level refinement. Experiments on a dataset with complex backgrounds and human models demonstrate the effectiveness of the proposed approach. In our future work, we will exploit the properties of texture consistency and symmetry of clothes to further improve the segmentation accuracy. In addition, failure cases are observed due to complex body posture and ambiguous boundaries between foreground and background. For these case, the superpixel cannot correctly match clothing edge, which is an age-old segmentation problem and is difficult to solve. We will explore pyramid superpixel segmentation with multiple granularities of superpixels to make up the drawback of single level of superpixels. Our ultimate goal is to propose unsupervised image segmentation algorithms which can efficiently and accurately extract clothing from images with fashion models and cluttered background. The ultimate goal is for web scale clothing image search.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (No. 61373121, No. 61036008), Program for Sichuan Provincial Science Fund for Distinguished Young Scholars (No. 2012JQ0029, No. 13QNJJ0149), Sichuan Science and Technology Innovation Seedling Fund (No. 2014-62), and the Fundamental Research Funds for the Central Universities (Project No. SWJTU09CX032, No. SWJTU10CX08).

References

[1] R. C. Gonzalez and R. E. Woods, "Digital imaging processing," *Massachusetts: Addison-Wesley*, 1992.

[2] I. Karoui, R. Fablet, J.-M. Boucher, and J. Augustin, "Variational region-based segmentation using multiple texture statistics," *IEEE Transactions on Image Processing*, vol. 19, no. 12, pp. 3146–3156, 2010.

[3] C. Rambabu, I. Chakrabarti, and A. Mahanta, "Flooding-based watershed algorithm and its prototype hardware architecture," in *IEE Proceedings-Vision, Image and Signal Processing*, vol. 151, no. 3, 2004, pp. 224–234.

[4] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[5] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.

[6] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *ACM Transactions on Graphics (TOG)*, vol. 23, no. 3, 2004, pp. 309–314.

[7] A. C. Gallagher and T. Chen, "Clothing cosegmentation for recognizing people," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, 2008, pp. 1–8.

[8] N. Wang and H. Ai, "Who blocks who: Simultaneous clothing segmentation for grouping images," in *Proc. of International Conference on Computer Vision (ICCV 2011)*, 2011, pp. 1535–1542.

[9] H. Chen, A. Gallagher, and B. Girod, "Describing clothing by semantic attributes," in *Proc. of Computer Vision (ECCV 2012)*, 2012, pp. 609–623.

[10] 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. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012)*, 2012, pp. 3330–3337.

[11] S. Liu, J. Feng, Z. Song, T. Zhang, H. Lu, C. Xu, and S. Yan, "Hi, magic closet, tell me what to wear!" in *Proc. of the 20th ACM international conference on Multimedia(ACM MM 2012)*, 2012, pp. 619–628.

[12] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012)*, 2012, pp. 3570–3577.

[13] X. Wu, L.-L. Liang, W.-J. Wang, and Q. Peng, "Principal object detection towards product image search," in *Proc. of International Conference on Audio, Language and Image Processing (ICALIP 2012)*, 2012, pp. 866 – 871.

[14] X. Wu, X.-P. Deng, L.-L. Liang, and Q. Peng, "Interactive product image search with complex scenes," in *Proc. of the 4th International Conference on Internet Multimedia Computing and Service (ICIMCS 2012)*, 2012, pp. 136–139.

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

[16] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[17] F.-H. Kong, "Image retrieval using both color and texture features," in *Proc. of International Conference on Machine Learning and Cybernetics (ICMLC 2009)*, vol. 4, 2009, pp. 2228–2232.

[18] J.-Q. Ma, "Content-based image retrieval with hsv color space and texture features," in *Proc. of International Conference on Web Information Systems and Mining (WISM 2009)*, 2009, pp. 61–63.

[19] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, 2011, pp. 409–416.