Clothing Extraction by Coarse Region Localization and Fine Foreground/Background Estimation

Xiao Wu, Bo Zhao, Ling-Ling Liang, and Qiang Peng

Department of Computer Science and Engineering, Southwest Jiaotong University
No. 111, North Section 1, 2nd Ring Road, Chengdu, China
{wuxiaohk,qpeng}@home.swjtu.edu.cn,
{zhaobo1987,eaily471956454}@gmail.com

Abstract. Online shopping is becoming more and more popular for billions of web users because of its convenience and efficiency. Customers can use content-based product image search engine to find their desired products. However, a frustrating fact is that the search results are significantly affected by the presence of natural backgrounds and fashion models. To minimize the influence of these noises, in this paper, an automatic clothing extraction algorithm is proposed, which consists of two phases: coarse clothing region localization with human proportion, and fine foreground/background modeling. Experiments on two datasets crawled from e-commerce websites demonstrate that the proposed approach achieves good performance, and has competitive performance with the interactive solution.

Keywords: Clothing Segmentation, Gaussian Mixture Model, Graph-based Image Segmentation, Foreground/Background Estimation.

1 Introduction

Nowadays, online clothing shopping becomes an attractive and convenient shopping way for millions of web users. Especially, with the emergence of social image sharing websites, such as *Pinterest*, it accelerates the progress of social and personalized e-commerce. There exist billions of diverse and beautiful clothes available on e-commerce websites, such as *Amazon*, *eBay*, and *Alibaba*. In order to attract the eyes of customers and demonstrate the actual appearance of clothes, the clothes are usually dressed by fashion models in real world and taken pictures with natural outdoor background. Therefore, a large portion of the apparel images in e-commerce websites commonly contain cluttered and complex backgrounds, which makes visual clothing search a challenging task.

Clothing segmentation and extraction, is an active research topic in computer vision and multimedia area. Its purpose is to identify and extract the clothing itself after removing the background and unrelated information. Existing clothing segmentation methods suffer from variations in colors and styles, different lighting conditions, geometric deformations, viewpoint changes, clustered backgrounds, and occlusions generated by poses or other objects. These variations are the major factors complicating the matters for clothing extraction.

S. Li et al. (Eds.): MMM 2013, Part II, LNCS 7733, pp. 316–326, 2013.

In this paper, we proposed an automatic clothing extraction algorithm by combining efficient graph-based image segmentation and foreground/background estimation. It mainly consists of two phases: a coarse clothing region localization and a fine clothing extraction. An apparel image is first segmented into multiple regions using a graph-based image segmentation approach. Skin and face regions are detected to guide the clothing region localization and assist the foreground/background model estimation. Based on the human proportion, inner and outer bound regions are roughly identified, indicating the potential clothing region and background region, respectively. Gaussian mixture model is adopted to build the foreground (clothing) and background models. By taking into account the spatial relationship among pixels, the generated GMM models are refined based on the components after efficient graph-based segmentation to achieve better segmentation performance. Experiments on two datasets crawled from e-commerce website *Taobao* and Pinterest like social sharing website *Mogujie* respectively demonstrate the proposed approach improves the segmentation results. It achieves competitive performance with the classic interactive segmentation approach GrabCut, from which users designate the desire region by dragging a rectangle around the object.

The rest of paper is organized as follows. Section 2 gives a brief overview of the related work. Section 3 elaborates the proposed clothing extraction algorithm. Section 4 presents the experiments. Finally, we summarize this paper with a conclusion.

2 Related Work

2.1 Product Image Search

In industry, shopping comparison website *Like.com*, is the first product image search engine to bring visual search for shopping, which builds an automated matching system for products, such as jewelries, handbags, shoes, and watches. It exploits computer vision and machine learning techniques to find similar-looking (similar colors, shapes, and patterns) products. In China, *Taotaosou* [13] under *Alibaba*, provides similar functions for visual product search. In academic research, *iLike* [3] explores vertical search by integrating textual and visual features to improve search performance, particularly targeting for product search of apparels and accessories. *iSearch* [10] combines global and local matching of local features to find similar product images in an interactive manner. A clothes search in consumer photos is presented in [15] by color matching and attribute learning, which leverages the low-level features (colors) and high-level features (attributes) of clothes. A Smart Mirror system [1] is proposed to recognize clothing styles and supports real-time fashion recommendation. However, the above-mentioned works mainly consider the images with clean background. The situation for product images with clustered background is not considered. To handle the discrepancies between online shopping images and daily photos, a two-step cross-scenario clothing retrieval is proposed via parts alignment and auxiliary set [11].

2.2 Clothing Segmentation

Image segmentation is widely used in many image related applications, such as content-based image retrieval, image annotation, and object recognition. In the past few decades, numerous image segmentation approaches have been proposed, including minimum spanning tree, min-cut, normalized cut, mean shift, and so on. Recently, a number of researches have been conducted on clothing image segmentation. Clothing modeling and recognition adopts an And-Or graph representation to produce a large set of composite graphical templates accounting for the wide variability of cloth configurations [2]. Without any pre-defined clothing model, a clothing segmentation method using foreground and background estimation is proposed [6]. A torso area is first detected based on dominant colors determination and then the background area is determined based on the Constrained Delaunay Triangulation (CDF). Using these two areas, the foreground and background estimation is obtained to accomplish the clothing segmentation task. However, in our work, we simply use the human proportion other than CDF to determine the foreground and background areas, which is more efficient. Given multiple images of the same person wearing the same clothing, the clothing co-segmentation [5] provides a significant improvement in recognition accuracy, by analyzing the mutual information between pixel locations near the face and the identity of the person to learn a global clothing mask. A multi-person clothing segmentation algorithm [14] is proposed for highly occluded images, which combines blocking models to address the person-wise occlusions.

3 Clothing Extraction

3.1 Framework

The presence of natural backgrounds and fashion models could significantly influence the performance of clothing image search. In order to identify the clothes in images and remove the impact of backgrounds and models, we proposed an automatic clothing extraction algorithm for clothing image database. The framework is illustrated in Fig. 1. It mainly consists of two phases: a coarse clothing region localization and a fine clothing extraction. To reduce the effect of noises in images, a Gaussian filter, as a preprocessing step, is first deployed to smooth the images. As skin and face are useful priori information to help locate the clothes, face and skin detection are adopted to detect the face and skin regions. According to the face region and human body proportions (face, torso, and so on), a coarse inner region and an outer region are identified, from which the potential clothing region and background region are roughly located. A fine-granularity clothing extraction is then undergone to accurately identify the clothes. To model the statistical distribution of image pixels, Gaussian Mixture Model (GMM) is adopted to build the foreground (clothing) and background models. At the same time, efficient graph-based image segmentation [4] is applied to segment the same image into multiple components, which act as an auxiliary resource. By taking into account the neighborhood and spatial relationship among pixels, the generated GMM models are refined to achieve better segmentation performance. Finally, the clothes are extracted from images, which can be used for visual clothing search to improve the performance.

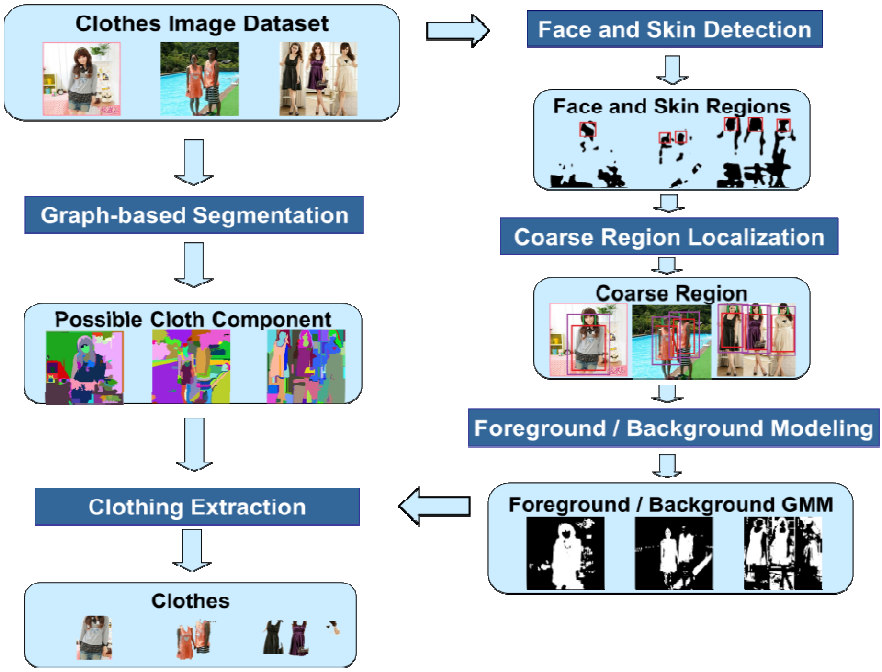

Fig. 1. Framework of the proposed clothing extraction

3.2 Skin and Face Detection

Our clothing extraction is guided by the detected faces. An Adaboost based face detector [7] is used in our work to locate the faces in different images, which has been widely used in many applications. It can accurately detect faces in real-life images with kinds of poses changes.

Skin pixels belong to the background, but sometimes some of them appear in the inner region and lie out of the determined background pixels. The wrong-classified skin pixels can affect the correct color distribution of both foreground and background, leading to poor segmentation. Skin pixels should be removed from the foreground seeds and be added into the background seeds to solve the above problem.

For skin detection, since Single Gaussian Model is sensitive to red-like pixels while Elliptical Boundary Model is sensitive to skin-like pixels [8], we combine Single Gaussian Model and Elliptical Boundary Model to obtain the skin area. A single Gaussian probability distribution using YCbCr color space is adopted to depict the skin distribution. Skin-color distribution is modeled through Gaussian joint probability distribution function. The parameters are estimated over all the color samples from the training data using Maximal Likelihood Estimation (MLE). The overlap of the results from Single Gaussian Model and Elliptical Boundary Model is treated as the final skin regions. The detected skin regions are illustrated in Fig. 2.

Fig. 2. The original images and the detected skin regions

3.3 Coarse Clothing Region Localization

As the clothing region is connected to the head, we exploit the detected face to guide the coarse clothing region localization. Though there are subtle differences between individuals, human proportions fit within a fairly standard range. In figure drawing, the basic unit of measurement is the "head", which is reasonably standard and has long been used to establish the proportions of the human figure. According to the study, an average person is generally 7-and-a-half heads tall (including the head) [16], which is shown in Fig 3(a).

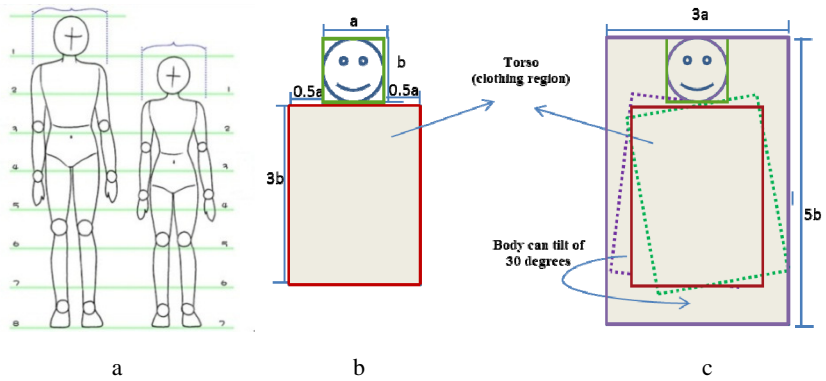

Fig. 3. Human proportions (a) and the inner bound (b) and outer bound (c)

A coarse clothing region can be firstly outlined based on the human proportions and clothing properties. Two rectangle regions called *inner bound* and *outer bound* are identified. The pixels in inner bound have high probability of belonging to the clothing, while the pixels outside the outer bound indicate the background. Assume that the width and height of the detected face are a and b, respectively, the region

positioned right below the face with the width and length ratio as 2a:3b is treated as the inner bound. The region including the face with the width and length ratio as 3a:5b is treated as the outer bound, which contains the face region and the inner bound region. The inner and outer bound are illustrated as red and purple rectangles in Fig. 3(b) and (c), respectively. Some examples with detected face, inner and outer bounds are shown in Fig. 4. These coarsely detected inner and outer bound will be exploited to construct the foreground and background models.

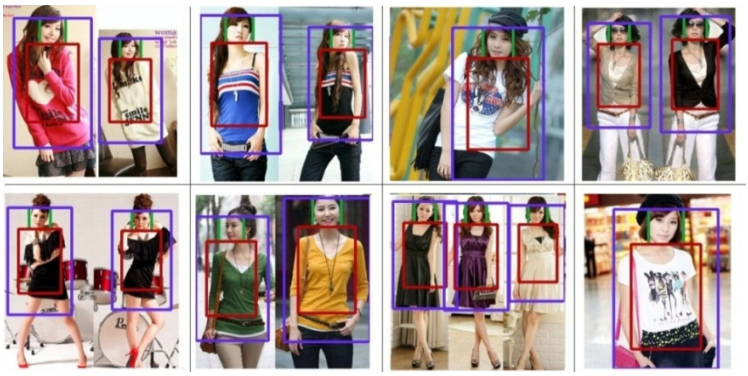

Fig. 4. The detected face, inner bound and outer bound regions

3.4 Clothing and Background Modeling

With the inner and outer bounds, the foreground (clothing) seeds are estimated from the inner region exclude the skin regions based on main colors determination, and the background (non-clothing) seeds are found based on the outer region plus skin regions. As foreground and background seeds contain several main colors, Gaussian Mixture Model (GMM) is employed to interpret color distributions of such mixture data.

Two GMMs are used to model the image color distributions of the clothing and background, respectively. In this work, the RGB color space is deployed.

$$\mathrm{p(x|clothes)}=\sum_{i=1}^{K_c}\pi_i^c\frac{1}{(2\pi)^{\frac{d}{2}}\left|\sum_i^c\right|^{\frac{d}{2}}}exp\left\{-\frac{1}{2}x-\mu_i^{cT}\sum_i^{c^{-1}}(x-\mu_i^c)\right\} \tag{1}$$

$$\mathrm{p(x|background)}=\sum_{\mathrm{i=1}}^{\mathrm{K_b}}\pi_{\mathrm{i}}^{\mathrm{b}}\frac{1}{(2\pi)^{\frac{\mathrm{d}}{2}}\left|\sum_i^b\right|^{\frac{\mathrm{d}}{2}}}exp\left\{-\frac{1}{2}x-\mu_i^{b^T}\sum_{\mathrm{i}}^{\mathrm{b}^{-1}}(x-\mu_i^b)\right\} \tag{2}$$

where x is a 3D vector standing for the RGB value of pixel x, μ_i^c and Σ_i^c are the mean value and covariance matrix of the *ith* Gaussian of the clothing GMM, μ_i^b and Σ_i^b are the mean value and covariance matrix of the *ith* Gaussian of the background

GMM. π_i^c and π_i^b are weighting factors of *ith* Gaussian of clothing and background respectively. All these parameters are determined by EM algorithm. K_c and K_b are the number of Gaussian distributions. In our experiments, they are set as 4.

GMM considers the statistical information which means pixels with similar color have the similar probability belongs to the clothing or background, but it ignores the spatial information which means pixels near each other should have similar probability. In addition, as GMM makes good use of the pixels' color properties, it is sensitive to illumination variations and clustered colors. To alleviate this problem, GMM-based color distribution integrates the efficient graph-based image segmentation [4] to improve the segmentation performance, which combines both the color properties and region properties.

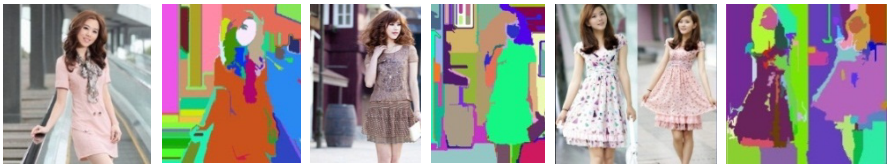

Fig. 5. Components after efficient graph-based image segmentation

To get space information, we consider the results of efficient graph-based segmentation [4] which cuts an image into several components. The detected components after efficient graph-based segmentation are shown in Fig. 5. For each component C_j after image segmentation, we calculate its foreground and background probabilities. The foreground probability $p(C_j|clothes)$ and background probability $p(C_j|background)$ are defined as the mean foreground probabilities and background probabilities of all pixels in the component, respectively, which are defined as follows:

$$\mathrm{p}(\mathrm{C}_j|\mathrm{clothes}) = \frac{1}{M}\sum_{i=1}^{M} p(x_i|clothes) \tag{3}$$

$$\mathrm{p}(\mathrm{C}_j|\mathrm{background}) = \frac{1}{M}\sum_{i=1}^{M} p(x_i|background) \tag{4}$$

where x_i is the *ith* pixel belongs to C_j and M is the total number of pixels in C_j.

The refined models $\mathrm{p}(x_i^j|\mathrm{clothes})$ and $\mathrm{p}(x_i^j|\mathrm{background})$ are determined by the combination of the original probability and the component probability. They consider the statistical information and spatial information, which are defined as:

$$\mathrm{p}(x_i^j|\mathrm{clothes}) = \mathrm{p}(x_i|\mathrm{clothes}) + \mathrm{p}(\mathrm{C}_j|\mathrm{clothes}) \tag{5}$$

$$\mathrm{p}(x_i^j|\mathrm{background}) = \mathrm{p}(x_i|\mathrm{background}) + \mathrm{p}(\mathrm{C}_j|\mathrm{background}) \tag{6}$$

The pixels are treated as the clothing pixel, if these pixels are within the outer bound region whose $\mathrm{p}(x_i^j|\mathrm{clothes}) > \mathrm{p}(x_i^j|\mathrm{background})$.

4 Experiments

There is no public product image dataset and corresponding ground truth available for evaluating the performance of clothing extraction. To evaluate the performance, we crawled product images from *Taobao*, the biggest e-commerce website in Asia. Totally, there are 1,356,901 images. These images are mainly from two categories: clothes and handbags. Since manually labeling the ground truth of clothing extraction on a dataset with millions of images is time-consuming, it is infeasible to evaluate on the whole dataset. We use two datasets: DS_TB, and DS_MGJ to evaluate the performance of the proposed solution. DS_TB consists of 1000 images with faces randomly selected from the above-mentioned dataset as the evaluation dataset. In addition, we crawled another 1000 clothing images from a Pinterest-like website in China, *Mogujie* (www.mogujie.com), which are mainly captured from outdoors, as the DS_MGJ.

Due to without the ground truth of accurate pixel-level clothing segmentation, it is impossible to evaluate the performance in an objective way. In this work, we use subjective evaluation for the performance of clothing extraction. Based on the clothing extraction results of different algorithms, five assessors were requested to evaluate the quality of clothing extraction by giving a score between 0 and 5 to the image, indicating the accuracy of the extracted clothing comparing to the perfect extraction. A higher score means a better segmentation performance. Score 5 refers to perfect clothing extraction, while 0 indicates that none of the extracted part belongs to the clothing. We use *average accuracy score* as the performance metric, which is defined as the sum of the scores for all images to the total number of images. In our work, N is 1000.

$$\text{aas} = \sum_{i=i}^{N} Score_i \,/\text{N} \tag{7}$$

To compare the performance, we compare the proposed solution with the Principal Object Detection (POD) [17], the simplified GMM based approach [6] and the interactive Grabcut [12]. Grabcut is an interactive image segmentation solution with human interaction by dragging a rectangle region in the query image to guide the object identification. Although the user interaction scheme is impractical for large scale object extraction, we evaluate the performance of the automatic solution compared to the interactive way. The principal object detection is induced from 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.

Fig. 6 demonstrates the average accuracy score of different approaches in datasets DS_TB and DS_MGJ. Overall, the proposed approach achieves the highest score compared with POD and GMM based approach in both datasets. In addition, without user interaction, the proposed solution has competitive performance as the interactive approach GrabCut. It means that our method can be applicable for large scale backend image datasets. The POD performs poor when facing images with complex backgrounds. Its performance is affected by the graph-based image segmentation. It

might select the wrong region as the principal object. It should be noted that the overall segmentation results in DS_MGJ are poorer than the ones in DS_TB. Most images in dataset DS_MGJ are captured outdoors with cluttered background, while parts of the images in DS_TB have relatively simple backgrounds. It makes the images in DS_MGJ are more challenging, which significantly affects the extraction performance. Fig. 7 shows the extracted results with different approaches. Generally, the extracted clothes using GrabCut and our method are more comprehensive and meaningful. Additionally, our algorithm is very efficient. On average, the clothing extraction process takes about 4 seconds per image on an Intel Core i5 3.1GHz processor with 4GBs of RAM.

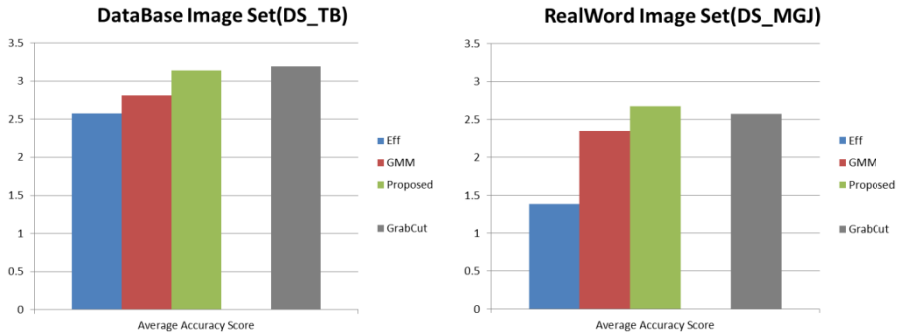

Fig. 6. Average accuracy score of different approaches in two datasets

Fig. 7. Performance comparison with different approaches

5 Conclusion

In this paper, we explore the clothing extraction algorithm with two steps: coarse clothing region localization and fine clothing extraction, which automatically localize the clothing region and estimate the foreground/background models to extract the clothing. Experiments on two datasets demonstrate the effectiveness of the proposed approach. In our future work, we will exploit the spatial symmetric property and texture consistency of clothes to further improve the segmentation accuracy. In addition, we will explore the clothing co-segmentation when there exist multiple images with similar clothes. Our ultimate goal is to propose unsupervised image segmentation algorithms which can efficiently and accurately extract clothing from images with cluttered background and fashion model.

Acknowledgements. The work described in this paper was supported by the National Natural Science Foundation of China (No. 61071184, 60972111, 61036008), Research Funds for the Doctoral Program of Higher Education of China (No. 20100184120009, 20120184110001), Program for Sichuan Provincial Science Fund for Distinguished Young Scholars (No. 2012JQ0029), the Fundamental Research Funds for the Central Universities (Project no. SWJTU09CX032, SWJTU10CX08, SWJTU11ZT08), and Open Project Program of the National Laboratory of Pattern Recognition (NLPR).

References

1. Chao, X., Huiskes, M.J., Gritti, T., Ciuhu, C.: A Framework for Robust Feature Selection for Real-time Fashion Style Recommendation. In: ICME, pp. 35–41 (2009)
2. Chen, H., Xu, Z., Liu, Z., Zhu, S.: Composite Templates for Cloth Modeling and Sketching. In: CVPR (2006)
3. Chen, Y., Yu, N., Luo, B., Chen, X.-W.: iLike: Integrating Visual and Textual Features for Vertical Search. In: ACM MM, pp. 221–230 (2010)
4. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient Graph-based Image Segmentation. IJCV 59(2), 167–181 (2004)
5. Gallagher, A.C., Chen, T.: Clothing Cosegmentation for Recognizing People. In: CVPR (2008)
6. He, Z., Yan, H., Lin, X.: Clothing Segmentation using Foreground and Background Estimation based on the Constrained Delaunay Triangulation. Pattern Recognition 41, 1581–1592 (2008)
7. Jones, M.J., Viola, P.: Fast Multi-view Face Detection. In: CVPR (2003)
8. Kakumanu, P., Makrogiannis, S., Bourbakis, N.: A Survey of Skin-color Modeling and Detection Methods. Pattern Recognition 40, 1106–1122 (2007)
9. Lee, J.Y., Yoo, S.I.: An Elliptical Boundary Model for Skin Color Detection. Science (2002)
10. Li, H., Wang, X., Tang, J.-H., Yi, L., Xiao, L.: iSearch: Towards Precise Retrieval of Item Image. In: ACM ICIMCS, pp. 5–8 (2011)
11. Liu, S., Song, Z., Liu, G., Xu, C.-S., Lu, H., Yan, S.-C.: Street-to-Shop: Cross-Scenario Clothing Retrieval via Parts Alignment and Auxiliary Set. In: CVPR (2012)

12. Rother, C., Kolmogorov, V., Blake, A.: Grabcut - Interactive Foreground Extraction using Iterated Graph Cuts. ACM SIGGRAPH 23, 309–314 (2004)
13. Taotaosou, http://www.taotaosou.com
14. Wang, N., Ai, H.: Who Blocks Who: Simultaneous Clothing Segmentation for Grouping Images. In: ICCV, pp. 1535–1542 (2011)
15. Wang, X., Zhang, T.: Clothes Search in Consumer Photos via Color Matching and Attribute Learning. In: ACM MM, pp. 1353–1356 (2011)
16. Wikipedia, http://en.wikipedia.org/wiki/Body_proportions
17. Wu, X., Liang, L.-L., Wang, W.-J., Peng, Q.: Principal Object Detection towards Product Image Search. In: ICALIP, pp. 866–871 (2012)