Learning non-target items for interesting clothes segmentation in fashion images

Costantino Grana, Simone Calderara, Daniele Borghesani, Rita Cucchiara
DII, University of Modena and Reggio Emilia
{*name.surname*}*@unimore.it*

Abstract

In this paper we propose a color-based approach for skin detection and interest garment selection aimed at an automatic segmentation of pieces of clothing. For both purposes, the color description is extracted by an iterative energy minimization approach and an automatic initialization strategy is proposed by learning geometric constraints and shape cues. Experiments confirms the good performance of this technique both in the context of skin removal and in the context of classification of garments.

1 Introduction

Generally, image segmentation is the process of partitioning the original image into different sub regions of homogeneity, which conveys saliency properties to be used in the following stages of image understanding.

In many applications segmentation is the first stage for target detection and selection, especially in large domain-specific datasets, where a class of target is depicted in different ways together with ancillary non-target and background objects. A typical context is that of e-commerce image datasets, and recently fashion goods have grown in popularity becoming a key market.

To handle automatic procedures of image retrieval, color classification, automatic tagging and annotation, clothing and fashion datasets often require a process of segmentation of the interesting garment. This step is mandatory, since usually each product is not pictured by itself but it is worn by a model, sometimes together with other complementary or decorative pieces of clothing, useful to improve its perception for marketing reasons. In this type of image datasets, normally the target object is in the central part of the image, surrounded by non-target elements that are a) the background; b) the support which can be either a mannequin or a human model; c) the ancillary fashion garments, such as skirts or part of trousers in case of a sweater, etc. Target fashion objects have the peculiarity to be variable in color and shape, often with different texture and hues so that uniform segmentation of target can become very hard. In fact, most segmentation algorithms [3] tackle the problem as a semi supervised learning problem, where a form of initialization is needed, but automatically selecting the starting points is the main challenge. Our statement is that in these conditions, learning and segmenting non-target objects is easier than learning and segmenting the target ones.

Figure 1. Model (a,b), Mannequin (c,d) and Still Life (e,f) phototype samples.

In this work we propose to learn what surely is not-target, by detecting skin and ancillary objects with an optimized classifier that exploits both fashion retailers photographic rules and an automatic phototype classification. Our proposal employs a Random Forest classifier and color based Gaussian Mixture Models (GMM) to select the piece of clothing of interest. Results are absolutely very promising and the procedure can be applied for other type of domain-dependent image datasets.

2 Related work

The literature on image segmentation is vast: please refer to works like [10, 5] for some recent surveys. Specifically, the use of segmentation for clothing can be considered another application scenario, in which the same techniques (based on supervised learning, clustering, and so on) are used not to describe the image given the parts, but to remove irrelevant regions, in order to focus on the interesting part (i.e. the currently advertised piece of clothing). Hu *et al.* [6] for ex-

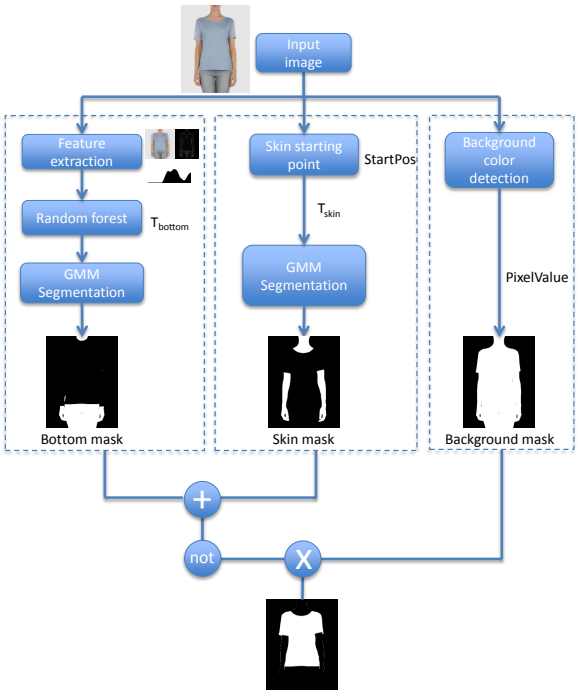

Figure 2. Block Diagram of the algorithm.

ample performed the segmentation via graph cuts with foreground and background seeds estimated by a constrained Delaunay triangulation. Bertelli *et al.* [1] integrated object-level top down information with low-level image cues into a kernelized structural SVM learning framework.

Most of the solutions to clothes selection tackle the problem under the hypothesis that clothes are worn by people, so finding where people is (for example with face detection) provides a good starting point for following steps. Skin represents one of the most valuable indicator of people presence and skin detection and removal is often adopted. When dealing with real fashion photo shootings and product advertising rules, most of these hypothesis are broken, because photographers aim to make the product appealing for the consumer, neglecting objective color reproduction and photo realism. This, indeed, constitutes a strong limit to trivially apply skin color identification on fashion photographs.

To deal with these problems adaptive skin detection approaches [9] have been used as an instrument to refine global skin detection results using the current image features. Among different skin color descriptors, Gaussian Mixture Models in color domain proved to be one of the state of the art approaches [7]. Since Gaussian Mixture Models training using the EM algorithm is computationally expensive, we follow the iterative energy minimization approach used in [8].

3 Segmenting target by segmenting non-target objects

We focus on large image datasets where the target object is depicted surrounded by similar non-target regions that are in general: *the background*, which is almost constant or often adjusted by photo retouching, and thus easily detectable; *the support* (e.g. a human model) which is very repetitive and thus can be inferred by the phototype; *ancillary non-target objects*, especially in the bottom part of the image, assuming that the target is acquired in the central part of the image.

Our system is composed of several modules as depicted in Fig. 2 and every single module will be detailed in the following subsections. Roughly, given an image, background removal is performed in order to obtain a binary mask. The *phototype detection* module first classifies the masks according to the shooting type (e.g. model is present in the image or mannequin is used). Consequently, according to the shooting type, both skin and additional garments and accessories are removed to obtain a clear picture of the object of interest.

3.1 Phototype identification

Different fashion products categories require various presentations in terms of photographic shooting to better highlight the product characteristics. Common fashion standards define three principal guidelines differentiating the presence of a human model wearing the garment (*Model* class), a mannequin dressed with the product (*Mannequin* class) and finally shoes and accessories are simply imaged without any distracting element (*Still Life* class). We will refer to this problem as *phototype identification* (see Fig. 1 for some visual samples).

An important aid in our work is that images have been previously subjected to photo retouching where shadows, minor skin flaws or possibly tattoos, and background are removed to provide a uniform appearance on the websites. So a simple processing of the images with a threshold based on the reference background color may be employed to recover the *binary mask* of the object. Images are further rescaled to a standard size.

Following the idea that the target is always at the center of the image, the information of the photo type is related to the shape of the object, and since we can avoid aligning the images, projection histograms (P_h, P_v) are a good candidate for this task. P_h is defined as $P_h(x|M) = \sum_{y=1}^{h}[M(x,y) == FG]$ where M is the $w \times h$ binary mask, $[\cdot]$ denotes the indicator function taking values $\{0,1\}$ and FG corresponds to the foreground value in the mask. The vertical projection P_v is defined accordingly.

Considering that, not all the bins are equally informative, we require to identify which elements are char-

acteristic of the different phototypes. For this reason we chose to use a discriminative approach which involves a feature selection process. The two principal solution employed in literature are Boosting and Decision Trees Classifiers. In particular Random Forest classifiers [2] have been chosen because they can handle easily multi-class problems providing an inherent features selection mechanism.

3.2 Human or non-human support segmentation

The segmentation approach is an iterative procedure which aims at minimizing the following Gibbs energy:

$$E(\alpha, k, \theta, z) = U(\alpha, k, \theta, z) + V(\alpha, z) \qquad (1)$$

where α is the current segmentation mask ($\alpha_n \in \{0, 1\}$), k is a vector, with $k_n \in \{1, \dots K\}$, assigning to each pixel a unique GMM component, one component either from the background or the foreground model. θ is the set of parameters of the GMM, and z is the image pixels. The data term $U(\cdot)$ is defined as

$$\begin{aligned} U(\alpha, k, \theta, z) = \sum_n & -log(p(z_n|\alpha_n, k_n, \theta)) \\ & - log(\pi(\alpha_n, k_n)) \end{aligned} \qquad (2)$$

where $p(\cdot)$ is a Gaussian probability distribution, and $\pi(\cdot)$ are mixture weighting coefficients. The smoothness term $V(\cdot)$ in Eq. 1 is given by:

$$V(\alpha, z) = \gamma \sum_{(m,n) \in C} [\alpha_n \neq \alpha_m] e^{-\beta \|z_m - z_n\|^2} \qquad (3)$$

where $[\cdot]$ denotes the indicator function taking values $\{0, 1\}$, and C is the set of pairs of neighboring pixels (8-way connectivity). This energy term encourages coherence in regions of similar colors. The minimization of Eq. 1 is efficiently performed by repeatedly estimating the GMM parameters θ and using minimum cut to estimate the segmentation mask.

The majority of approaches for segmentation in images, mitigate the problem of initialization (of what is foreground to keep and what is background to remove), assuming some level of human interaction. Even if this turns out to be an effective solution, especially in the application scenarios for which these algorithms usually are designed (i.e. assistance for the image post-processing in studio), when a fully automated system is desired and no process changes in the company's production flow are possible, an automatic initialization becomes necessary.

To solve this problem, we can consider that in most fashion datasets for internet e-commerce applications, models faces are located in the top-center of the image, so it is safe to expect that by selecting part of the object mask from the top, skin tones and hair will be identified. Following these photographic guidelines it is feasible to assume that the *skin mask* can be initialized by naively selecting few lines from the top of the image, proportionally w.r.t. image size, use this element as the initializers to solve the optimization problem of Eq. 1 and then retain the generated mask.

3.3 Non-target detection

Conversely, it is not equally safe to hypothesize that, by selecting part of the object mask from the bottom, some items and parts of the image which are not relevant to define the main clothing under interest (legs potentially left out by the skin segmentation, shoes, trousers and so on) will be identified. In fact the presentation is conceived to advertise the specific product type: sometimes full shot of the body are taken (e.g. long dresses), while in shirts and jackets a zoomed shooting is preferred to highlight garments details. This elements concur to raise the degree of variability of the bottom part of the image, and fixed geometric constraints (i.e. choosing a fixed number of lines from the bottom) typically lead to under/over detection.

We face this problem as a supervised learning task where the irrelevant lower part (*bottom mask*) is directly inferred combining image features and operators experience. Using a regression random forest classifier we obtain the number of probably irrelevant lines from the bottom of the image. The features used are again the projection histograms of the segmentation mask, along with two low resolution maps of the colors and of the gradients. These two maps help to distinguish cases in which the silhouette by itself is not sufficient, for example cases with boots, leggings, transparencies and so on. The use of redundant features is mitigated by the feature selection stage of the random forest classifier that chooses the best ones for the assigned task. In particular, color and gradient maps have non-zero feature weights in most of the cases. As for the skin mask case, we solve the optimization of Eq. 1 using the learnt bottom lines for initialization to obtain a binary mask of the non-informative bottom part.

The garment selection is finally obtained by the binary AND of the *binary mask* and the complement of the binary OR combination of the two aforementioned masks.

4 Experimental Results

In order to verify the correctness of our proposal we built, in collaboration with a worldwide leader in fashion e-commerce, an annotated dataset for garment of interest selection and segmentation. The dataset is composed of 60204 images of clothes from most famous fashion designers, divided into the three phototype cat-

Table 1. Phototype classification confusion matrix.

	Model	Mannequin	Still Life
Model	99,63%	0,35%	0,02%
Mannequin	0,00%	99,62%	0,38%
Still Life	0,15%	0,70%	99,15%

egories, mentioned in Sec. 3.1, where 23% are mannequin shootings, 52% models and the remaining 25% still life. Results have been visually checked by fashion operators that judged as either acceptable or not the automatically selected image region containing the garment of interest.

Table 1 reports the confusion matrix for the phototype classification task, which demonstrates the remarkable accuracy of the algorithm.

Table 2 shows the accuracy results of our proposal, comparing the skin detection module with the recent state of the art technique in [4]; performance have been evaluated both globally and separately for every single step to underline the solid performance of the proposed algorithms. Fig. 3 shows the output of the proposed garment selection method, capable of segmenting even complex pieces of clothes like highly textured dresses or shawls. Most of the mistakes are due to erroneous skin detection or to the presence of pieces of clothes that exhibit strong chromatic characteristic affine to skin color tones, but our proposal strongly mitigates them w.r.t. conventional skin detection algorithms as can be deduced from the comparison in Table 2.

Figure 3. Results on garment selection. Red is the skin mask and Green the bottom mask.

Table 2. Garment selection results.

		skin	bottom	sel.acc.
Model	[4]	78.2	96.6	76.8
	our	**93.1**	96.6	**92.3**
Mannequin		n.a.	98.5	98.5
Full dataset	[4]	n.a.	97.2	83.5
	our	n.a.	97.2	**94.2**

5 Conclusions

We proposed a complete system for garment selection which has the great advantages of being adaptable to different fashion rules and precise enough to compete with human operators performance. The use of learning techniques allows to reconfigure the system rules by simply providing new examples to the different classifiers. We are currently testing the system in the industrial workflow of a world leader e-commerce retailer.

References

[1] L. Bertelli, T. Yu, D. Vu, and B. Gokturk. Kernelized structural svm learning for supervised object segmentation. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 2153–2160, 2011.

[2] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[3] J. Carreira and C. Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 3241–3248, June 2010.

[4] A. Cheddad, J. Condell, K. Curran, and P. Mc Kevitt. A skin tone detection algorithm for an adaptive approach to steganography. *Journal of Signal Processing*, 89(12):2465–2478, Dec. 2009.

[5] J. Freixenet, X. Muñoz, D. Raba, J. Martí, and X. Cufí. Yet another survey on image segmentation: Region and boundary information integration. In *European Conference on Computer Vision*, volume 2352, pages 408–422, 2002.

[6] Z. Hu, H. Yan, and X. Lin. Clothing segmentation using foreground and background estimation based on the constrained delaunay triangulation. *Pattern Recognition*, 41:1581–1592, May 2008.

[7] P. Kakumanu, S. Makrogiannis, and N. G. Bourbakis. A survey of skin-color modeling and detection methods. *Pattern Recognition*, 40(3):1106–1122, 2007.

[8] C. Rother, V. Kolmogorov, and A. Blake. "GrabCut": interactive foreground extraction using iterated graph cuts. In *ACM SIGGRAPH*, pages 309–314, 2004.

[9] H.-M. Sun. Skin detection for single images using dynamic skin color modeling. *Pattern Recognition*, 43(4):1413–1420, 2010.

[10] H. Zhang, J. E. Fritts, and S. A. Goldman. Image segmentation evaluation: A survey of unsupervised methods. *Computer Vision and Image Understanding*, 110(2):260–280, 2008.