Real-Time Semantic Clothing Segmentation

George A Cushen and Mark S Nixon

University of Southampton, UK
{gc505, msn}@ecs.soton.ac.uk

Abstract. Clothing segmentation is a challenging field of research which is rapidly gaining attention. This paper presents a system for semantic segmentation of primarily monochromatic clothing and printed/stitched textures in single images or live video. This is especially appealing to emerging augmented reality applications such as retexturing sports players' shirts with localized adverts or statistics in TV/internet broadcasting. We initialise points on the upper body clothing by body fiducials rather than by applying distance metrics to a detected face. This helps prevent segmentation of the skin rather than clothing. We take advantage of hue and intensity histograms incorporating spatial priors to develop an efficient segmentation method. Evaluated against ground truth on a dataset of 100 people, mostly in groups, the accuracy has an average F-score of 0.97 with an approach which can be over 88% more efficient than the state of the art.

1 Introduction

Clothing can be considered to be one of the core cues of human appearance and segmentation is one of the most critical tasks in image processing and computer vision. Clothing segmentation is a challenging field of research which few papers have addressed. It can benefit augmented reality [1], human detection [2], recognition for re-identification [3], pose estimation [4], and image retrieval for internet shopping. Although the field has recently been gaining more attention, a real-time clothing segmentation system remains challenging. This is primarily due to the wide diversity of clothing designs, uncontrolled scene lighting, dynamic backgrounds, variation in human pose, and self and third-party occlusions. Secondly, difficult sub-problems such as face detection are usually involved to initialize the segmentation procedure.

In this paper, we present a clothing segmentation method for single images and video which can segment upper body clothing of multiple persons in real-time, as summarized in Figure 1. Our approach is primarily designed to benefit emerging augmented reality applications [1]. These include computer gaming and augmenting localized adverts or statistics onto players' shirts for close-up shots in live TV/internet broadcasting. Shirts worn by sports teams, such as in major league basketball, are often uniformly coloured with text and a logo to indicate the player and team. For this reason, we focus on the case of predominantly monochromatic tops, and attempt to additionally segment any textures on them which can be useful for the purpose of retexturing.

Our main contributions include:

1. An efficient automated method for accurate segmentation of multiple persons wearing primarily uniformly coloured upper body clothing which may contain textured regions. Spatial priors are employed and each set of resulting cloth and texture contours are semantically labelled as such and associated to a face. Unlike most previous work which evaluates visually or with respect to applications (such as recognition), we evaluate the segmentation directly and quantitatively against a dataset of 100 people.
2. An initialization scheme where initial points on the cloth are located by estimating skin colour and employing an iterative colour similarity metric to locate the clothing. This can prevent initializing cloth points on the skin in the case of clothing with deep neck lines such as vests and many female tops.

Previous work is described in Section 2. An efficient approach for segmenting clothing is presented in Section 3. Performance is quantitatively assessed in Section 4, and conclusions are drawn in Section 5.

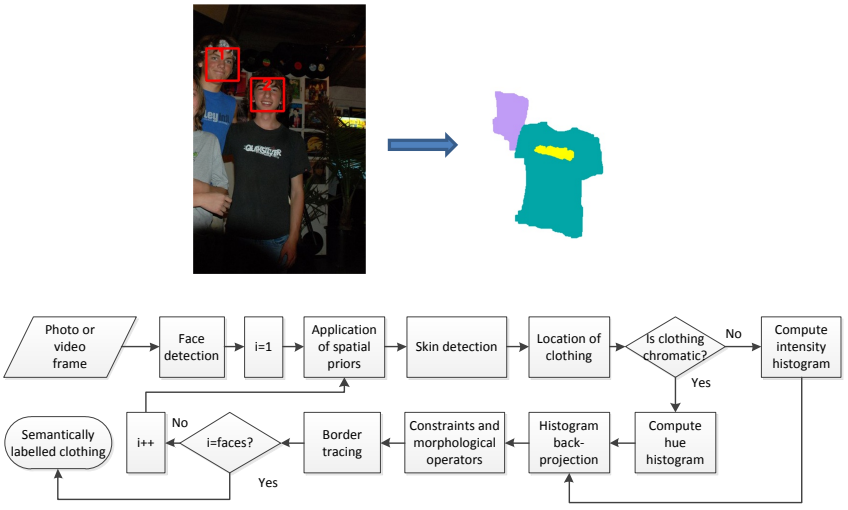


Fig. 1. Segmenting clothing with texture in groups of people. Top: the coloured clothing segments on the right, with texture (logos) depicted in yellow, correspond to the numerically labelled persons on the left. Bottom: system diagram.

2 Previous Work

One of the most popular approaches to clothing segmentation involves a Markov Random Field (MRF) framework based on graph cuts [5, 6, 3]. Although these

approaches have robustness to a diverse range of clothing, they can suffer in accuracy, producing very crude segmentation. This is especially true in cases of occlusions and difficult poses. The MRF has since been reformulated to deal with groups of people [7]. More recently, Wang and Ai [9] introduced a clothing shape model which is learned using Random Forests and self-similarity features, with a blocking model to address person-wise occlusions. The fastest approaches in literature are our previous histogram work [1] (12fps for the overall multithreaded application) and the region growing approach presented in [8]. Although [8] extracts the person including skin pixels, their approach is fast, reporting 16.5ms per detected person for segmenting clothing and 10fps overall (including face detection and a classification application). Their private dataset was captured in a controlled lab setup, featuring a predominantly white background.

Face detection is generally employed in existing approaches to initialize points on the cloth. It should be noted that due to frontal face detection, the segmentation approaches are limited to frontal poses. The initial cloth points are located by applying a (scaled) distance from the bottom of the detected face. In the case of clothing with deep neck lines, such as vests and many female tops, these methods can segment the skin rather than the cloth. In contrast, we design a more complex initialization scheme which attempts to avoid this.

The majority of previous work focusses on segmenting a single image of a single person offline. Contrary to these methods, we attempt to simultaneously process multiple persons and maintain reasonable accuracy whilst increasing computational efficiency to enable real-time image/video processing. Furthermore, no existing approaches yield a semantically labelled segmentation which includes contours for any textured regions, such as logos, on the cloth.

3 Clothing Segmentation

3.1 Pre-Processing and Initialization

For pre-processing, the single image or video frame is converted from RGB to the more intuitive and perceptually relevant HSV colour-space. The corresponding illumination channel is then normalized, giving image N. This helps to alleviate, to some extent, the non-uniform effects of uncontrolled scene lighting. Additionally, a 3×3 box blur is performed as a simple denoising measure, yielding image I. We let the H, S, and V channels correspond to I_0, I_1, and I_2 respectively and use the OpenCV HSV intervals $I_0 = [0, 180]$ and $I_{1,2} = [0, 255]$. For our image notation, we also refer to the origin as the top left of the image.

A chromatic/achromatic mask is defined where achromatic pixels are those with illumination extremes or low saturations:

$$\texttt{chrome}(I) = 0 \leq I_0 \leq 180 \wedge 26 \leq I_1 \leq 255 \wedge 26 \leq I_2 \leq 230 \tag{1}$$

Viola-Jones face detection is performed on image N as a prerequisite for our segmentation approach. This technique is based on a cascade architecture for reasonably fast and accurate classification with OpenCV's popular frontal face

trained classifier cascade. We limit the region of interest for object detection to the top half of the image in order to further increase efficiency. For each face detected, the segmentation procedure in the following sections is performed.

3.2 Spatial Priors

To increase robustness against hues/intensities in the background which are similar to those on the clothing, and to increase computational efficiency, we constrain segmentation of each person to a region of interest (ROI). The size of this region is determined by detecting faces in our training dataset (see Section 4) and studying the upper body clothing bounds, given by anatomy and pose, relative to the detected face size and position. As a result of these studies, spatial priors are defined as 5 times the detected face height and 4.5 times the face width and positioned as follows:

$$\begin{aligned}\texttt{crop}(I) = \texttt{Rect}(&\texttt{Point}(F_x - 1.75F_{width}, F_y + 0.75F_{height}),\\ &\texttt{Point}(F_x + 2.75F_{width}, F_y + 0.75F_{height} + 5F_{height}))\end{aligned} \tag{2}$$

where the F vector for each person is output by face detection. The bounds of the ROI are also clipped to within the image dimensions.

3.3 Locating Points on the Clothing

Points on the clothing are required in order to initialise segmentation. Previous work often employs a scaled distance from a detected face to achieve this. However, this approach is susceptible to initialising clothing points on the skin in the case of clothing with deep neck lines such as vests and many female tops, and hence the segmentation has reduced accuracy. We propose a solution to this problem. The faces detected on the training dataset in the previous section are scaled to within 80×80 pixels, whilst maintaining their aspect ratios. We study the average face and define a region which tends to primarily be skin pixels and avoids occlusion by long hair:

$$\texttt{FSkin}(I) = \texttt{Rect}(\texttt{Point}(F_x + 15s, F_y + 36s), \texttt{Point}(F_x + 65s, F_y + 56s)) \tag{3}$$

where the scale factor $s = F_{width}/80$. The skin colour α is estimated by computing the mean of the pixels in the $\texttt{FSkin}(I)$ region.

A sparse iterative procedure is established across the $x = [F_x, F_x + F_{width}]$ and $y = [F_y + F_{height}, F_y + 2F_{height}]$ intervals, shifting a 5×5 pixel window. During each iteration, the mean colour β of the window is computed. The HSV colour similarity between the window's mean β and the estimated skin colour α is calculated. The two cylindrical HSV colour vectors are transformed to Euclidean space using the following formulae:

$$x = \cos(2I_0) \cdot I_1/255 \cdot I_2/255, \;\; y = \sin(2I_0) \cdot I_1/255 \cdot I_2/255, \;\; z = I_2/255 \tag{4}$$

The Euclidean distance d is then computed between the 3D colour points. If $d \leq 0.35$, we assume the window primarily contains skin pixels. The bottom of the clothing's neck, $Neck_y$, is located as the lowest 'skin window' within the aforementioned x and y intervals. Note that in the case that the subject is wearing clothing which is so similar in colour to their skin that the colour similarity distance remains below the threshold, we establish a cloth sampling window located around the x-coordinate of the face centre at the end of the y-interval. Otherwise, in the typical case, the cloth sampling window is located beneath the garment's neck at:

$$\begin{aligned} \texttt{sample}(I) = &\texttt{Rect}(\texttt{Point}(F_x + 0.25F_{width}, Neck_y + 1.5\gamma), \\ &\texttt{Point}(F_x + 0.75F_{width}, Neck_y + 1.5\gamma + 0.25F_{height})) \end{aligned} \tag{5}$$

where γ refers to the aforementioned window size of 5 pixels.

3.4 Chromatic vs Achromatic

We design a histogram based approach because this is very efficient and can have a high accuracy on segmenting clothing which is primarily monochromatic. In such cases, it can also be suitable for semantic segmentation of printed/stitched textures within the clothing. First, we determine the chromatic ratio of the clothing which is estimated by taking the mean of the binary image $\texttt{chrome}(I)$ (see Equation 1) with the sampling ROI of Equation 5 applied:

$$\text{Chromatic Ratio} = r = \frac{1}{0.5F_{width} \cdot 0.25F_{height}} \sum_{x,y \in \texttt{sample}(I)} \texttt{chrome}(I(x,y)) \tag{6}$$

Second, the image plane for segmentation is determined based on whether the clothing is primarily achromatic or chromatic:

$$\text{Segmentation Plane} = S = \begin{cases} I_0 & \text{if } r \geq 0.5 \\ I_2 & \text{otherwise} \end{cases} \tag{7}$$

Based on these two cases, we empirically define some segmentation parameters in Table 1:

Table 1. Segmentation Parameters

	Segmentation Plane S	
Parameter	Hue I_0	Intensity I_2
q	16	15
λ	50	3

3.5 Clothing Segmentation

This section describes our histogram based segmentation routine. A histogram $\{g\}_{i=1\dots q}$ is computed for image plane S with the ROI $\texttt{sample}(I)$ applied:

$$g_i = \sum_{x,y \in \texttt{sample}(I)} \delta[b(x,y) - i]. \tag{8}$$

where δ is the Dirac delta function and let $b\colon \mathbb{R}^2 \to \{1 \dots q\}$ be the function which maps the pixel at location $S(x,y)$ to the histogram bin index $b(x,y)$. We empirically choose to quantize to q bins as this provides a good compromise between under-segmentation (due to variation in cloth hue/intensity caused by lighting) and over-segmentation (due to objects with similar hues/intensities which are in direct contact with the clothing). Quantization reduces the computational and space complexity for analysis, clustering similar color values together. The histogram is then normalized to the discrete range of image intensities:

$$h_i = \min\left(\frac{255}{\max(g)} \cdot g_i, 255\right), \forall i \in 1 \dots q \tag{9}$$

where h is the normalized histogram, g is the initial histogram, and subscripts denote the bin index.

Image S is back-projected to associate the pixel values in the image with the value of the corresponding histogram bin, generating a probability distribution image P where the value of each pixel characterizes the likelihood of it belonging to the clothing (i.e. histogram h). The resulting probability image is thresholded to create a binary image:

$$P(x,y) = \begin{cases} 255 & \text{if } P(x,y) \geq \lambda \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Scene conditions such as illumination can alter the perceived hue/intensity of the cloth, so we empirically set the λ threshold relatively low (see Table 1).

We further constrain P by considering $\texttt{chrome}(I)$, the computed chromatic mask. If $S = I_0$, we let $P = P \wedge \texttt{chrome}(I)$. Otherwise, if $S = I_2$ and $r \leq 0.05$, we constrain with the achromatic mask, letting $P = P \wedge (255 - \texttt{chrome}(I))$. In the unlikely case that the sampled clothing pixels are mostly achromatic but not entirely (i.e. for $0.05 < r < 50$), we do not constrain P with the achromatic mask as it can exhibit significant holes at the location of coloured cloth pixels.

Morphological closing with a kernel size of 3×3 is employed to remove small holes, followed by opening, with the same kernel, to remove small objects. Experimentally, this has been found to fill small holes in the edges of the clothing caused by harsh lighting, and remove small objects of a similar hue/intensity to the cloth which are in contact with it from the camera's perspective.

Suzuki-Abe border tracing is employed to extract a set T of contours with corresponding tree hierarchy H. We choose to limit the hierarchy to 3 levels deep as this can provide sufficient information for the clothing contour, potential contours for a printed/stitched texture within, and potential holes within the

texture contour(s). The top level of the hierarchy is iterated over, computing the bounding box area of each contour. The area is approximated to that of the bounding box for efficiency and experimentally this appears to be acceptable. We define the largest contour T_{max} as the clothing. Therefore, there is robustness to objects in the scene which have a similar hue/intensity as the clothing but are not in contact with it from the camera's perspective. An initial clothing segmentation mask $\tilde{M}$ can be defined by filling T_{max}.

The initial clothing segment $\tilde{M}$ can suffer in accuracy in cases of harsh illumination or patterned clothing. Robustness to these cases can be increased by sequentially performing morphological closing and opening with a large kernel size. The reason for not employing this larger kernel on the first iteration of morphological operations is that this can decrease accuracy if it is not just the clothing segment of interest present, but also other large objects of similar hue/intensity in the background. Since the morphological processes can create additional contours, border tracing is computed again to extract the clothing as one segment M.

3.6 Texture Segmentation

Existing clothing segmentation methods do not purposely attempt to semantically segment printed/stitched textures within clothing masks. Segmentation of any potential printed designs on clothing can be used to make the clothing cue more informative. We hypothesize that this could be useful for the purpose of re-texturing in emerging augmented reality clothing applications.

We iterate through the contours T in the second level of the contour hierarchy H, computing their areas. Unlike the area computation for the cloth contour, we do not approximate by bounding boxes here because textures can have more variation in shape, which may result in inaccurate area estimations. We consider contours with areas above a dynamic empirically defined threshold of $0.25F_{width}F_{height}$ to belong to a printed texture on the clothing. If no contour above the threshold is found, then we assume that there is no texture. Otherwise the extracted contours are filled and the regions of their corresponding hole contours in the third level of the hierarchy are subtracted (if they exist) from this result, yielding the texture mask.

4 Experiments

We study the quality of the clothing segmentation, robustness to noise, and the computational timing. Results are reported in Table 2, alongside a comparison with the state of the art. Two datasets are combined: Soton [10] and Images of Groups [11]. The Soton dataset consists of images of individuals whereas the Images of Groups dataset is very challenging for segmentation, featuring real-world Flickr photos of groups. A testing subset of 100 persons is formed from images featuring predominantly uniformly coloured upper body clothing and does not feature groups where clothing of adjacent persons is of a very similar

colour and in direct contact. A training subset of 50 persons is randomly selected from remaining images in the dataset for use in Sections 3.2 and 3.3.

Table 2. Histograms can provide efficient and effective clothing segmentation. The accuracy values are for indicative purposes only as they are not directly comparable.

	Our Method	[8]	[9]	[3]
Timing	**2.0**ms per person on 2.93GHz core	16.5ms per person on 3.16GHz core	Offline	Offline
Accuracy	**0.97** F-score	N/A	92.8%	89.4%

To compare the computational efficiency of our approach to the closest state of the art, we similarly consider static image regions for each person with resolution 200×300. Our approach, excluding pre-processing (Section 3.1), achieves on average 2ms per person using a 2.93GHz CPU core. Thus our method is over 88% more efficient than results reported by [8] under similar conditions, with the exception that they employ a faster CPU core (3.16GHz). Our overall system, including pre-processing (face detection and simple denoising), is fast, achieving results at an average rate of 25fps (frames per second) for segmenting one person given an input resolution of 480×640 pixels. The equivalent computation time is dissected as 38ms pre-processing per image and 2ms clothing segmentation per person. Face detection is our biggest computational bottleneck, so for high resolutions, the input to face detection could be downscaled. The segmentation procedure could easily be parallelized for each face detected, if using a multi-core CPU and the average number of persons to segment justifies the threading overhead.

Accuracy is reported using the best F-score criterion: $F = 2RP/(P + R)$, where P and R are the precision and recall of pixels in the cloth segment relative to our manually segmented ground truth. We achieve an average F-score over the entire testing dataset of 0.97. Since the F-score reaches its best value at 1 and worst at 0, our approach shows good accuracy. Additionally, by visual inspection of Figures 1 and 2, we can see that our approach can semantically segment clothing of persons in various difficult uncontrolled scenes with some robustness to minor occlusions (Figure 2(c)) and minor patterns (Figure 2(b)). Clothing segmentation literature tends to report accuracy with regards to applications (such as recognition or classification) rather than directly on segmentation. Although not directly comparable, the performance is higher than that reported in [9], using mostly images from the same dataset.

Finally, we consider robustness to one of the most common forms of noise: additive white Gaussian noise. This is caused by random fluctuations in the pixels. Naturally, this could be easily filtered but our aim is to demonstrate robustness. If the input image is represented by I_{input}, and the Gaussian noise by Z, then we can model a noisy image by simply adding the noise: $I_{noisy} = I_{input} + Z$. Z consists of 3 planes which correspond to the RGB planes of I_{input}, and is drawn from a zero-mean normal distribution with standard deviation σ. We

study the effects of noise on a randomly selected image of a single person. Figure 3(a) depicts a graph of accuracy (F-score) versus the noise standard deviation (255σ) which shows our approach can handle significant noise. Note that we multiply σ by 255 since we consider integer images, and there is no data point plotted for $\sigma = 0.9$ because face detection mistakenly detects two faces and thus there are two results. Noise can positively affect our segmentation, for example at $\sigma = 0.2$, if noise pixels with hues similar to the cloth are established in dark clothing regions which originally had many unstable hues. At $\sigma = 1.0$, depicted by Figure 3(b), the face detection accuracy continues to decrease; however, the corresponding clothing segmentation in Figure 3(c) remains reasonably accurate. The segmentation fails entirely at $\sigma = 1.1$ since the prerequisite of face detection fails to detect any faces.

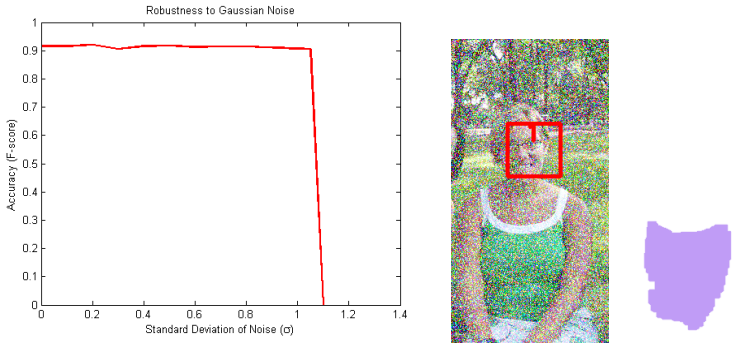
Our approach is subject to some limitations. We assume that there are no significant objects of a similar hue/intensity to the chromatic/achromatic clothing which are in direct contact with it from the camera's perspective, and the clothing is predominantly uniformly coloured (i.e. it is not significantly patterned). These limitations should not significantly affect the suggested computer games and broadcasting applications for the purpose of augmented reality.

(a) (b) (c)

Fig. 2. Further segmentation results. Each pair shows the numerically labelled person(s) on the left with their corresponding colour labelled clothing on the right.

5 Conclusions

We have presented an algorithm for automatic semantic clothing segmentation of multiple persons. It does so by estimating whether the clothing is chromatic or achromatic and then applying a histogram based approach on the hues or intensities. In order to initialise points on the cloth, we have proposed a method consisting of skin colour estimation and colour similarity to locate the bottom of the garment's neck. We have shown that the proposed framework is able to segment clothing more efficiently than existing state of the art methods, whilst achieving good accuracy and robustness on a difficult dataset. Although our approach is limited to predominantly uniformly coloured clothing (which may



(a) (b) (c)

Fig. 3. Robustness to noise: (a) graph of accuracy versus Gaussian noise σ, (b) input with considerable noise ($\sigma = 1.0$), and (c) corresponding clothing segmentation.

contain textured regions), it may be of particular benefit to emerging real-time augmented reality applications such as sports broadcasting and computer gaming.

References

1. Cushen, G., Nixon, M.: Markerless Real-Time garment retexturing from monocular 3D reconstruction. In: IEEE ICSIPA, Malaysia (2011) 88–93
2. Sivic, J., Zitnick, C.L., Szeliski, R.: Finding people in repeated shots of the same scene. In: BMVC. Volume 3. (2006) 909–918
3. Gallagher, A.C., Chen, T.: Clothing cosegmentation for recognizing people. In: CVPR 2008, IEEE (2008) 1–8
4. Lee, M.W., Cohen, I.: A model-based approach for estimating human 3D poses in static images. IEEE TPAMI (2006) 905–916
5. Schnitman, Y., Caspi, Y., Cohen-Or, D., Lischinski, D.: Inducing semantic segmentation from an example. In: ACCV 2006, Springer (2006) 373–384
6. Hu, Z., Yan, H., Lin, X.: Clothing segmentation using foreground and background estimation based on the constrained Delaunay triangulation. Pattern Recognition **41** (2008) 1581–1592
7. Hasan, B., Hogg, D.: Segmentation using Deformable Spatial Priors with Application to Clothing. In: BMVC. (2010) 1–11
8. Yang, M., Yu, K.: Real-time clothing recognition in surveillance videos. In: IEEE ICIP. (2011) 2937–2940
9. Wang, N., Ai, H.: Who Blocks Who: Simultaneous Clothing Segmentation for Grouping Images. In: ICCV. (2011)
10. Seely, R.D., Samangooei, S., Lee, M., Carter, J.N., Nixon, M.S.: The University of Southampton Multi-Biometric Tunnel and introducing a novel 3D gait dataset. In: BTAS, IEEE (2008) 1–6
11. Gallagher, A., Chen, T.: Understanding Images of Groups Of People. In: CVPR. (2009)