ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

A Robust Human Body Segmentation Based on Bottom-Up Approaches

Pantrangi Sindhu Gowthami
PG Scholar,
Department of ECE,
S.O.E.T,
Sri Padmavathi Mahila Viswa Vidyalayam,
Tirupati.

Kallapalli Suseela
Assistant Professor
Department of ECE,
S.O.E.T,
Sri Padmavathi Mahila Viswa Vidyalayam,
Tirupati.

ABSTRACT

Segmentation of human bodies in images is a challenging task that can facilitate numerous applications, like scene understanding and activity recognition. In order to cope with the highly dimensional pose space, scene complexity, and various human appearances, the majority of existing works require computationally complex training and template matching processes. We propose a bottom-up methodology for automatic extraction of human bodies from single images, in the case of almost upright poses in cluttered environments. The position, dimensions, and color of the face are used for the localization of the human body, construction of the models for the upper and lower body according to anthropometric constraints, and estimation of the skin color.

Different levels of segmentation granularity are combined to extract the pose with highest potential. The segments that belong to the human body arise through the joint estimation of the foreground and background during the body part search phases, which alleviates the need for exact shape matching. The performance of our algorithm is measured using 40 images (43 persons) from the INRIA person dataset and 163 images from the "lab1" dataset, where the measured accuracies are 89.53% and 97.68%, respectively. Qualitative and quantitative experimental results demonstrate that our methodology outperforms state-of-the-art interactive and hybrid top-down/bottom-up approaches.

Index Terms: ***Adaptive skin detection, anthropometric constraints, human body segmentation, and multilevel image segmentation.***

INTRODUCTIONS

Extraction of the human body in unconstrained still images is challenging due to several factors, including shading, image noise, occlusions, background clutter, the high degree of human body deformability, and the unrestricted positions due to in and out of the image plane rotations. Knowledge about the human body region can benefit various tasks, such as determination of the human layout, recognition of actions from static images, and sign language recognition. Human body segmentation and silhouette extraction have been a common practice when videos are available in controlled environments, where background information is available, and motion can aid the segmentation through background subtraction. In static images, however, there are no such cues, and the problem of silhouette extraction is much more challenging, especially when we are considering complex cases.Moreover, methodologies that are able to work at a frame level can also work for sequences of frames, and facilitate existing methods for action recognition based on silhouette features and body skeletonization.

In this study, we propose a bottom-up approach for human body segmentation in static images. We decompose the problem into three sequential problems: Face detection, upper body extraction, and lower body extraction, since there is a direct pairwise correlation among them. Face detection provides a strong indication about the presence of humans in an image, greatly reduces the search space for the upper body, and provides information about skin color. Face dimensions also aid in determining the dimensions of the rest of the body, according to anthropometric constraints.

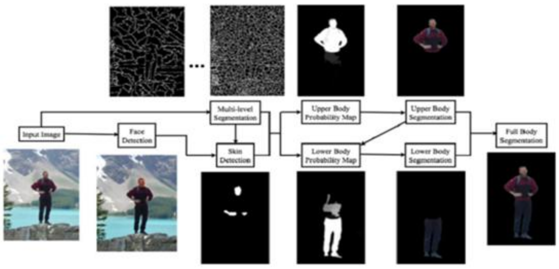

Fig.1: Overview of the methodology. Face detection guides estimation of anthropometric constraints and appearance of skin, while image segmentation provides the image's structural blocks.

The regions with the best probability of belonging to the upper body are selected and the ones that belong to the lower body follow.

The search for the upper body, which in turns leads the search for the lower body. Moreover, upper body extraction provides additional information about the position of the hands, the detection of which is very important for several applications. The basic units upon which calculations are performed are super pixels from multiple levels of image segmentation. The benefit of this approach is twofold. First, different perceptual groupings reveal more meaningful relations among pixels and a higher, however, abstract semantic representation. Second, a noise at the pixel level is suppressed and the region statistics allow for more efficient and robust computations. Instead of relying on pose estimation as an initial step or making strict pose assumptions, we enforce soft anthropometric constraints to both search a generic pose space and guide the body segmentation process. An important principle is that body regions should be comprised by segments that appear strongly inside the hypothesized body regions and weakly in the corresponding background. The general flow of the methodology can be seen in Fig. 1.

The major contributions of this study address upright and not occluded poses:
1) We propose a novel framework for automatic segmentation of human bodies in single images.
2) We combine information gathered from different levels of image segmentation, which allows efficient and robust computations upon groups of pixels that are perceptually correlated.
3) Soft anthropometric constraints permeate the whole process and uncover body regions.
4) Without making any assumptions about the foreground and background, except for the assumptions that sleeves are of similar color to the torso region, and the lower part of the pants is similar to the upper part of the pants, we structure our searching and extraction algorithm based on the premise that colors in body regions appear strongly inside these regions (foreground) and weakly outside (background).

PROPOSED METHOD

FACE DETECTION

Localization of the face region in our method is performed using Open CV's implementation of the Viola–Jones algorithm that achieves both high performance and speed. The algorithm utilizes the AdaBoost method on combinations of a vast pool of Haar-like features, which essentially aim in capturing the underlying structure of a human face, regardless of skin color. Since skin probability in our methodology is learned from the face region adaptively, we prefer an algorithm that is based on structural features of the face.

The Viola–Jones face detector is prone to false positive detections that can lead to unnecessary activations of our algorithm and faulty skin detections. To refine the results of the algorithm, we propose using the skin detection method presented, and the face detection algorithm presented in. The skin detection method is based on color constancy and a multilayer perception neural network trained on images collected under various illumination conditions both indoor and outdoor, and containing skin colors of different ethnic groups. The face detection method is based on facial feature detection and localization using low-level image processing techniques, image segmentation, and graph-based verification of the facial structure.

First, the pixels that correspond to skin are detected using the method. Then, the elliptical regions of the detected faces in the image found by the Viola–Jones algorithm are evaluated according to the probabilities of the inscribed pixels. More specifically, the average skin probability of the pixels X of potential face region FRi , for each person i, is compared with threshold T Global Skin (set empirically to 0.7 in our experiments). If it passes the global skin test (greater than T Global Skin), it is further evaluated by our face detector. If the facial features are detected, then FRi is considered to be a true positive detection.

After fitting an ellipse in the face region, we are able to define the fundamental unit with respect to which locations and sizes of human body parts are estimated, according to anthropometric constraints. This unit is referred to as palm length (PL), because the major axis of the ellipse is almost the same size as the distance from the base of the palm to the tip of the middle finger. Thus, our anthropometric model is adaptive for each person and invariant to scale.

MULTIPLE-LEVEL IMAGE SEGMENTATION

Relying solely on independent pixels for complicated inference leads to propagation of errors to the high levels of image processing in complex real-world scenarios. There are several different sources of noise, such as the digital sensors that captured the image, compression, or even the complexity of the image itself and their effect is more severe at the pixel level. A common practice to alleviate the noise dwelling at the pixel level is the use of filters and algorithms that extract collective information from pixels. Moreover, groups of pixels express higher semantics. Small groups preserve detail and large groups tend to capture shape and more abstract structures better. Finally, computations based on super pixels are more efficient and facilitate more flexible algorithms.

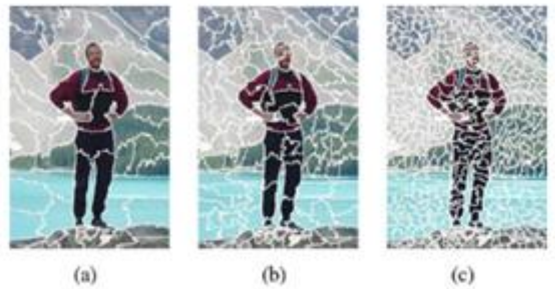

Fig. 2.Image segmentation for 100, 200, and 500 super pixels.

In this study, we propose using an image segmentation method, in order to process pixels in more meaningful groups. However, there are numerous image segmentation algorithms, and the selection of an appropriate one was based on the following criteria. First, we require the algorithm to be able to preserve strong edges in the image, because they are a good indication of boundaries between semantically different regions. Second, another desirable attribute is the production of segments with relatively uniform sizes. Studies on image segmentation methods show that although these algorithms approach the problem in different ways, in general, they utilize low-level image cues and, thus, their results cannot guarantee compliance with the various and subjective human interpretations. Thus, we deem this step as a high-level filtering process and prefer to oversegment the image; therefore, as not to lose detail. Region size uniformity is important because it restrains the algorithm from being tricked by over segmenting local image patches of high entropy (e.g., complex and high detailed textures) at the expense of more homogeneous regions that could be falsely merged, although they belong to semantically different objects (e.g., human hand over a wooden surface with color similar to skin). use super pixels instead of single pixels and normalized cuts for segmenting the image. The method we adopt in this study is the entropy rate super pixel segmentation (ERSS) algorithm, which provides a tradeoff between accuracy and computational complexity.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

This approach is based on optimizing an objective function consisting of two components: The entropy rate of a random walk on a graph and a balancing term. Results of the ERSS are shown in Fig. 2. More importantly, we propose using multiple levels of segmentation, in order to alleviate the need for selecting an appropriate number for the regions to be created and combine information emanating from different perceptual groupings of pixels. Although our framework can accept any number of segmentation levels, we find that two segmentation levels of 100 and 200 segments provide accurate results.

For the skin detection algorithm, a finer segmentation of 500 super pixels is used, because it manages to discriminate better between adjacent skin and skin-like regions, and recover skin segments that are often smaller compared with the rest image regions.

Skin Detection:

Among the most prominent obstacles to detecting skin regions in images and video are the skin tone variations due to illumination and ethnicity, skin-like regions and the fact that limbs often do not contain enough contextual information to discriminate them easily. In this study, we propose combining the global detection technique with an appearance model created for each face, to better adapt to the corresponding human's skin color (Fig. 5.2). The appearance model provides strong discrimination between skin and skin-like pixels, and segmentation cues are used to create regions of uncertainty.

Regions of certainty and uncertainty comprise a map that guides the Grab Cut algorithm, which in turn outputs the final skin regions. False positives are eliminated using anthropometric constraints and body connectivity. An overview of the process can be seen in Fig. 3. Each face region FR is used to construct an adaptive color model for each person's skin color. In this study, we propose using the r, g, s, I, Cr, and a.

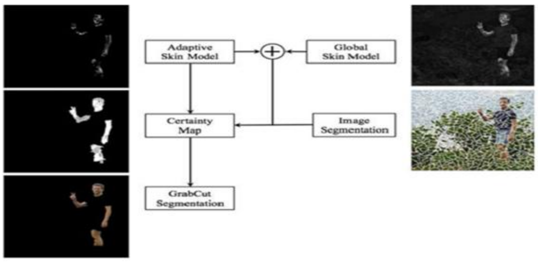

Fig. 3. Skin detection algorithm

channels. In more detail, r = R/(R + G + B), g = G/(R + G + B), and s = (R + G + B)/3; therefore, r and g are the normalized versions of the R and G channels, respectively, and s is used instead of b to achieve channel independence. Channels I, Cr, and a from YIQ (or NTSC), YCbCr, and Lab colors paces, respectively, are chosen because skin color is accentuated in them. The skin color model for each person is estimated after fitting a normal distribution to each channel, using the pixels in each FR . The parameters that represent the model are the mean values μij and standard deviations σ ij for each FR and channel j = 1 ... 6 for channels r, g, s, I, Cr, and a. Each image pixel's probability of being a skin pixel is calculated separately for each channel according to a normal probability distribution with the corresponding parameters. We expect true skin pixels to have strong probability response in all of the selected channels. The skin probability for each pixel X is as follows:

$$\boldsymbol{P_{Skin_i}(X) = \prod_{j=1}^{6} \mathcal{N}(X, \mu_{ij}, \sigma_{ij})} \qquad \textbf{(1)}$$

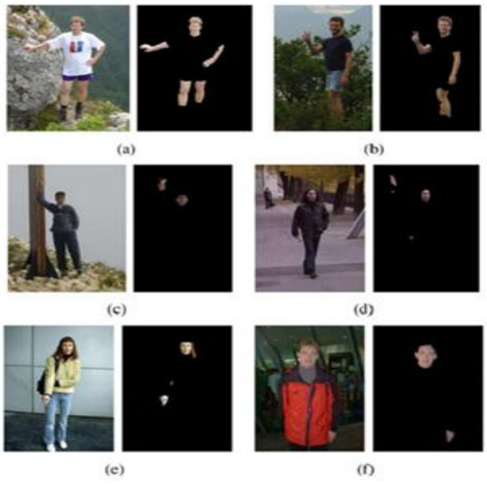

Fig.4. Skin detection examples

The adaptive model in general focuses on achieving a high score of true positive cases. However, most of the time it is too "strict" and suppresses the values of many skin and skin-like pixels that deviate from the true values according to the derived probability distribution. At this point, we find that an influence of the skin global detection algorithm is beneficial because it aids in recovering the uncertain areas. Another reason we choose to extend the skin detection process is that relying solely on an appropriate color space to detect skin pixels is often not sufficient for real-world applications. The two proposals are combined through weighted averaging (with a weight of 0.25 for the global model, and 0.75 for the adaptive model). The finest level of image segmentation is used at this point to characterize segments as certain and probable background and foreground. For the certain foreground regions, however, only the pixels with sufficiently high probability in the adaptive model are used as seeds; therefore, as to control their strong influence. In order to characterize a region as probable background or foreground, its mean probability of the combined probability must be above a certain threshold (empirically set to 0.2 and 0.3, respectively). Examples can be seen in Fig. 5.

Upper Body Segmentation:

In this section, we present a methodology for extraction of the whole upper human body in single images, extending [40], which dealt with the case, where the torso is almost upright and facing the camera. The only training needed is for the initial step of the process, namely the face detection and a small training set for the global skin detection process. The rest of the methodology is mostly appearance based and relies on the assumption that there is a connection between the human body parts. Processing using super pixels instead of single pixels, which are acquired by an image segmentation algorithm, yield more accurate results and allow more efficient computations.

The initial and most crucial step in our methodology is the detection of the face region, which guides the rest of the process. The information extracted in this step is significant. First, the color of the skin in a person's face can be used to match the rest of his or her visible skin areas, making the skin detection process adaptive to each person. Second, the location of the face provides a strong cue about the rough location of the torso. Here, we deal with cases, where the torso is below the face region, but without strong assumptions about in and out of plane rotations. Third, the size of the face region can further lead to the estimation of the size of body parts according to anthropometric constraints. Face detection here is primarily conducted using the Viola–Jones face detection algorithm for both frontal and side views. Since face detection is the cornerstone of our methodology, we refine the results of the aforementioned method using the face detection algorithm presented.

Once the elliptical region of the face is known, we proceed to the foreground (upper body) probability estimation. To better utilize the existing spatial and color relations of the image pixels, we perform multiple level oversegmentation and examine the resulting superpixels. We regard superpixels with color similar to that of the face region as skin and superpixels with color similar to the regions inside torso masks as clothes. With respect to clothes, the size of face's ellipse guides the construction of rectangular masks for the foreground using anthropometric constraints. Our basic assumption is that a good foreground mask should contain regions that appear mostly inside the mask and not outside (background). In other words, we try to identify "islands of saliency," in the aforementioned sense. As opposed to approaches based on pose estimation, we employ simple heuristics to conduct a fast and rough torso pose estimation and guide the segmentation process.
The torso is usually the most visible body part, connected to the face region and in most cases below it. Using anthropometric constraints, one can roughly estimate the size of the torso and its location. However, different poses and head motion make torso localization a challenging task, especially when assumptions about poses are relaxed. Instead of searching for the exact torso region or using complex

pose estimation methods, we propose using a rough approximation of the torso mask in order to identify the most concentrated island of saliency. This criterion allows for fast inference about the torso's size and location, while relieving the need for the complex task of explicit torso estimation, without sacrificing accuracy.

As discussed, different levels of segmentation give rise to different perceptual pixel groupings, and each segment is described by the statistics of its color distribution. In each segmentation level,each segment is compared with the rest and its similarity image is created, depicting the probabilistic similarity of each pixel to the segment. Similarly to the skin detection process, normal probability distributions according to the mean μi and standard deviation σi of segment Si are estimated for each channel j = 1, 2, 3 of the Lab color space, and the probability for each image pixel belonging to this probability is calculated. We estimate the final probability as the product of the probabilities.

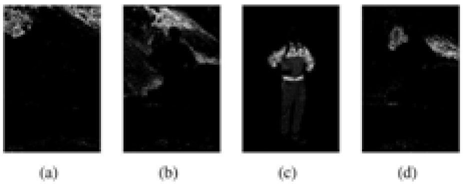

Fig.5: Example of similarity images for random segments.

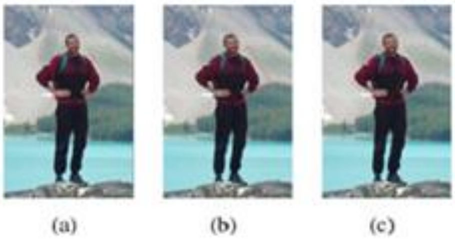

Fig.6:Masks used for torso localization.

In Each Channel Separately

Example similarity images are shown in Fig. 5. The resulting image that depicts the probability segment Si that is the same color as the rest of the segments is referred to as the similarity image. Similarity images are gathered for all of the different segmentation levels l. Here, we use two segmentation levels in this stage of 100 and 200 super pixels, because they provide a good tradeoff between perceptual grouping and computational complexity.

$$P_{SimIm_{li}}(X) = \prod_{j=1}^{3} \mathcal{N}(X, \mu_{ij}, \sigma_{ij}) \qquad (2)$$

Sequentially, a searching phase takes place, where a loose torso mask is used for sampling and rating of regions according to their probability of belonging to the torso. Since we assume that sleeves are more similar to the torso colors than the background, this process combined with skin detection actually leads to upper body probability estimation. The mask is used for sufficient sampling instead of torso fitting; therefore, it is estimated as a large square with sides of 2.5PL, with the top most side centered with respect to the face's center. In order to relax the assumptions about the position and pose of the torso, the mask is rotated by 30 ◦ left and right of its initial position (0 ◦) (see in Fig. 5.5). By using a large square mask and allowing this degree of freedom, we manage to sample a large area of potential torso locations. By constraining its size according to anthropometric constraints, we make the foreground/background hypotheses more meaningful.

During the search process, the mask is applied to each similarity image and its corresponding segment is scored. Let Torso Mask be a binary image, where pixels are set to 1 (or "on") inside the square mask and 0 (or "off") outside so that Simile ∩ Torso Mask selects the probabilities of the similarity image that appear inside the mask. Index t = 1, 2, 3 corresponds to a torso mask at angle −30, 0, or 30. Thus, (3) and (4) rate each segment's potential of belonging to the foreground and background, respectively, and (5) combines the two potentials in the form of a ratio as follows:

$$P_{FG}(S_{tli}) = \sum^{|S_{tli}|} \mathrm{SimIm}_{li} \cap \mathrm{TorsoMask}_t \qquad (3)$$

$$P_{BG}(S_{tli}) = \sum^{|S_{tli}|} \mathrm{SimIm}_{li} \cap \overline{\mathrm{TorsoMask}_t} \qquad (4)$$

$$\mathrm{TorsoScore}(S_{tli}) = \frac{P_{FG}(S_{tli})}{P_{BG}(S_{tli}) + \epsilon}. \qquad (5)$$

we can achieve accurate and robust results without imposing computational strain. The obvious step is to threshold the aggregated potential torso images in order to retrieve the upper body mask.

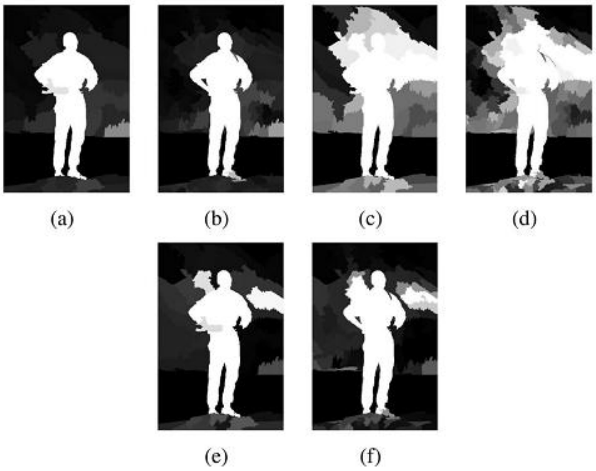

Fig.7. Segments with potential of belonging to torso. (a), (b) For segmentation level 1 and 2 and torso mask at 0 ◦. (c), (d) For segmentation level 1 and 2 and torso mask at 30 ◦. (e) (f) For segmentation level 1 and 2 and torso mask at −30 ◦.

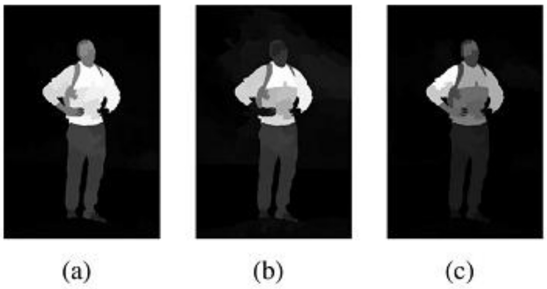

Fig.8.Aggregation of torso potentials shown in Fig. 7, for torso masks at 0 ◦, 30 ◦, and −30 ◦.

In most cases, hands or arms' skin is not sampled enough during the torso searching process, especially in the cases, where arms are outstretched. Thus, we use the skin masks estimated during the skin detection process, which are more accurate than in the case they were retrieved during this process, since they were calculated using the face's skin color,in a color space more appropriate for skin and segments created at a finer level of segmentation. These segments are superimposed on the aggregated potential torso images and receive the highest potential.

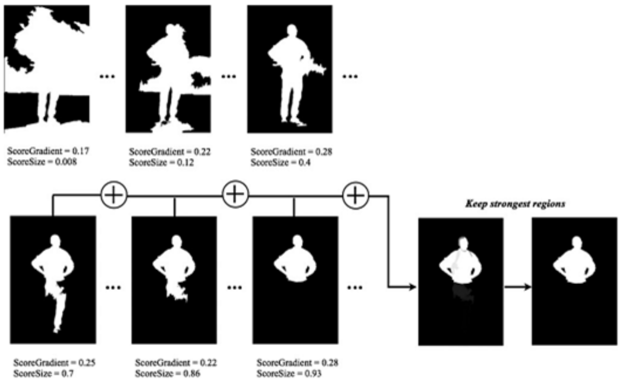

Fig.9. Thresholding of the aggregated potential torso images and final upper body mask. Note that the masks in the top row are discarded.

Instead of using a simple or even adaptive thresholding, we use a multiple level thresholding torecover the regions with strong potential according to the method described, but at the same time comply with the following criteria:

1) they form a region size close to the expected torso size (actually bigger in order to allow for the case, where arms are outstretched), and

2) the outer perimeter of this region overlaps with sufficiently high gradients. The distance of the selected region at threshold t (Region) to the expected upper body size (Exp Upper Body Size) is calculated as follows:

$$\textbf{ScoreSize} = \mathbf{e}^{\frac{-|\mathrm{Region}_t - \mathrm{ExpUpperBodySize}|}{\mathrm{ExpUpperBodySize}}} \qquad \textbf{(6)}$$

whereExp Upper Body Size = 11 × PL2 . The score for the second criterion is calculated by averaging the gradient image (Grad Im) responses for the pixels that belong to the perimeter (Region) of Region as

$$\textbf{Score Grad} = \frac{1}{|PRegion_t|} \qquad \textbf{(7)}$$

Thresholding starts with zero and becomes increasingly stricter at small steps (0.02). In each thresholding level, the largest connected component is rated, and the masks with Score Grad > 0.05 and Score Size > 0.6 are accumulated to a refined potential image (see in Fig. 5.8). Incorporation of this a priori knowledge to the thresholding process aids the accentuation of the true upper body regions (UBR). Accumulation of surviving masks starts when Score Size > 0.6 and resulting masks after this point will keep getting closer monotonically to the expected region size. Accumulation ends when Score Size drops below 0.6. The rationale behind this process is to both restrict and define the thresholdingrange and focus the interest to segments with high potential of forming the upper body segment. The aggregate mask (Aggregate Mask) can now be processed easily and produce more meaningful results. Specifically, we set a final threshold, which allows only regions that have survived more than 20% of the accumulation process in the final mask for the UBR. This process is performed for every initial torso hypothesis; therefore, in the end, there are three corresponding aggregate masks, out of which the one that overlaps the most with the initial torso mask and obtains the highest aggregation score is selected. The aggregation score shows how many times each pixel has appeared in the accumulation process, implicitly implying its potential of belonging to the true upper body segment.

Refinement

In many cases, the extracted upper body mask is very accurate and can be used as a final result. However, we choose to add an extra refinement step to cope with probable segmentation errors and pixels that manage to survive the multiple thresholding process. One idea that we use here is to give the upper body mask as input to an interactive foreground/background algorithm that requires “seeds” corresponding to the foreground and background. Grow Cut and Grab Cut are used for experiments.

Grow Cut expects the RGB image as input and a map denoting the seeds for background, foreground, and uncertain pixels, whereas Grab Cut can operate on a more refined map containing the certain foreground, certain background, probable foreground, and probable background regions. In order to construct these maps, we employ morphological operations on the upper body mask, with adaptive square structural elements (SEs) according to anthropometric constraints. For GrowCut, the uncertain region is constructed by dilating the upper body mask with a SE with sides equal to PL/6, the face’s ellipse with a SE with sides equal to PL/10 and the skin regions with a SE with sides equal to PL/12. Possible holes between the face and torso region are also filled. The certain foreground is similarly constructed with erosion instead of dilation, where the sides of the SEs are now PL/4, PL/4, and PL/10, respectively. The rest of the map is classified as background. For the Grab Cut algorithm, the possible background ground is constructed by dilating the upper body mask, the face’s ellipse and skin masks using SEs with sides PL/10, PL/2, and PL/12, the probable foreground is constructed by eroding the masks with SEs with sides PL/4, PL/4, and PL/10, respectively, and the certain foreground by eroding them with SEs with sides PL/1, PL/3, and PL/8, respectively. Both algorithms are guided by the extracted upper body mask; therefore, their results are very similar. Their main difference is that Grab Cut can make better guesses in cases of uncertainty and segment large regions loosely defined by the map, whereas GrowCut is more sensitive to the map and more influenced by background seeds. In Fig. 10, for example, both algorithms extract the upper body successfully, but Grow Cut removes the small enclosed regions by the arms, whereas Grab Cut includes them.

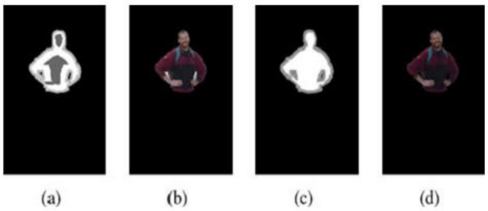

Fig.10. Example of foreground/ background certainty maps and segmentations for (a), (b) Grab Cut and (c), (d) Grow Cut.

Lower Body Extraction:

The algorithm for estimating the lower body part, in order to achieve full body segmentation is very similar to the one for upper body extraction. The difference is the anchor points that initiate the leg searching process. In the case of upper body segmentation, it was the position of the face that aided the estimation of the upper body location. In the case of lower body segmentation, it is the upper body that aids the estimation of the lower body’s position. More specifically, the general criterion we employ is that the upper parts of the legs should be underneath and near the torso region. Although the previously estimated UBR provides a solid starting point for the leg localization, different types of clothing like long coats, dresses, or color similarities between the clothes of the upper and lower body might make the torso region appear different (usually longer) than it should be. To better estimate the torso region, we perform a more refined torso fitting process, which does not require extensive computations, since the already estimated shape provides a very good guide.

The expected dimensions of the torso are again calculated based on anthropometric constraints, but in a more accurate model. In addition, in order to cope with slight body deformations, we allow the rectangle to be constructed according to a constrained parameter space of highest granularity and dimensionality. Specifically, we allow rotations with respect to rectangle's center by angle φ, translations in x- and y-axes, τx and τ y and scaling in x- and y-axes, sx and sy . The initial dimensions of the rectangle correspond to the expected torso in full frontal and upright view and it is decreased during searching in order to accommodate other poses. The rationale behind the fitting score of each rectangle is measuring how much it covers the UBR, since the torso is the largest semantic region of the upper body, defined by potential upper body coverage (UBC), while at the same time covering less of the background region, defined by potential S (for Solidity). Finally, in many cases, the rectangle needs to be realigned with respect to the face's center (FaceCenter) to recover from misalignments caused by different poses and errors. A helpful criterion is the maximum distance of the rectangle's upper corners (LShoulder, RShoulder) from the face's center (Dsf), which should be constrained. Thus, fitting of the torso rectangle is formulated as a maximization problem.

$$\boldsymbol{\theta max f(\theta) = \alpha_1 \times UBC(\theta) + \alpha_2 \times S(\theta) + \alpha_2 \times D_{s\,f}(\theta)} \qquad \textbf{(8)}$$

Where $\boldsymbol{\theta = (\phi, T_x, T_y, s_x, s_y)}$

$$\mathbf{UBC}(\boldsymbol{\theta})=\frac{\boldsymbol{\Sigma TorsoMask(\theta) \cap UBR}}{\boldsymbol{\Sigma TorsoMask(\theta)}}$$

$$\mathbf{S}(\boldsymbol{\theta})=\frac{\boldsymbol{\Sigma TorsoMask(\theta)}}{\boldsymbol{\Sigma UBR}}$$

$$\boldsymbol{D_{s\,f}(\theta) = e^{\frac{-\left|Max_{D_{sf-1.5\times}}{}^{PL}\right|}{1.5\times PL}}}$$

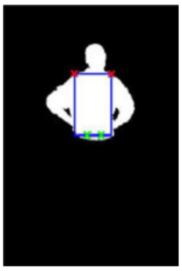

Fig.11. Best torso rectangle with shoulder and beginning of the legs positions.

we estimate the shoulder positions (top corners of the rectangle), and more importantly, the waist positions (lower corners of the rectangle). In turn, waist positions approximately indicate the beginning of the right and left leg legBR = (x,y) and legBL = (x,y), respectively. These points are the middle points of the line segments of the waist points and the point in the center of the line that connects them. Fig. 11 shows a case of a fitted torso and the aforementioned points. Similarly to upper body extraction and the torso rectangle fitting case, we explore hypotheses about the leg positions using rectangles by first creating rectangle masks for the upper leg parts and using them as samples for the pants color and finally perform appearance matching and evaluate the result. The assumption we make here is that there is uniformity in the color of the upper and lower parts of the pants.In the case of short pants, where the lower leg parts are naked, the previously calculated skin regions are used to recover them. In order to reduce computational complexity, the size and position of the upper leg rectangles are fixed and adhering to anthropometric constraints and the only free parameter is their angle of rotation with respect to their center φright and φleft . Let Leg Mask(θ) be the binary mask for the two hypothesized leg parts, where θ = (φright,φleft). Every possible upper leg mask is used as a sample of the pants regions, and the leg regions are estimated using the clothes and skin detection process (1)–(5). An example mask can be seen in Fig. 12. The hypothesized foreground is the pixels that belong to the leg mask, and background is the rest of the image plus the pixels of the upper body mask, without the pixels below the waist line segment (if any). The leg mask retrieved from each hypothesis is the largest connected component of image segments with color similar to the hypothesis and the skin regions retrieved in the previous steps. There is no strong need for precise alignment of the masks and the real leg parts, just enough coverage is pursued in order to perform a useful sampling. Thus, the algorithm can recover from slight torso misalignment and performs well in cases of different leg positions, without imposing the computational strain of dense searching using dense mask parameters.

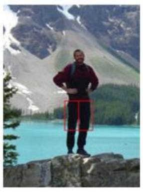

Fig.12. Example legs mask for φright = 0 and φleft = 0.

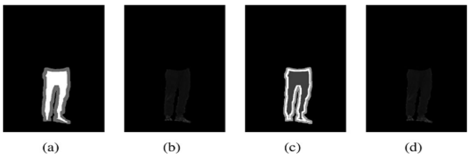

Fig.13. Example of foreground/background certainty maps and segmentations for (a) and (b) Grab Cut and (c) and (d) Grow Cut.

After the leg potentials are found, the same thresholding process as in the case of the upper body takes place, with the difference that now the expected lower body size is used in (6) (Exp Lower Body Size instead of Exp Upper Body Size), where Exp Lower Body Size = 6 × PL2 . In order to construct the tri map of Grow Cut to perform the refinement process for the leg regions, the leg mask is eroded by a square structuring element (SE) with side PL/4 followed by dilation by a SE with side PL/5 in order to create the uncertainty mask, and for the certain foreground mask it is eroded using a SE with side PL/3. Fig. 13 shows an example. In some cases, thin and ambiguous regions like belts or straps might end up belonging to both the upper and lower body, or in the worst case the background.

Most of the time, however, the refinement of the upper and lower regions is able to recover them, and during merging of the two regions they are included in the final outcome.

RESULTS

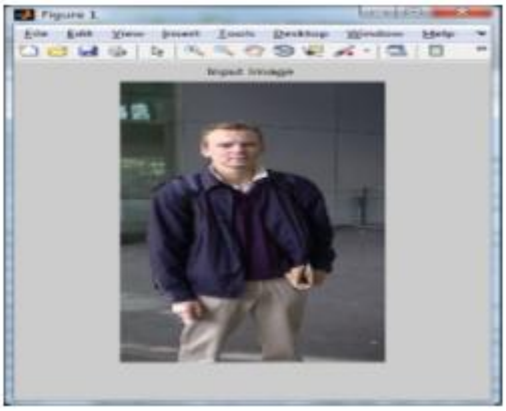

Fig.14. input image

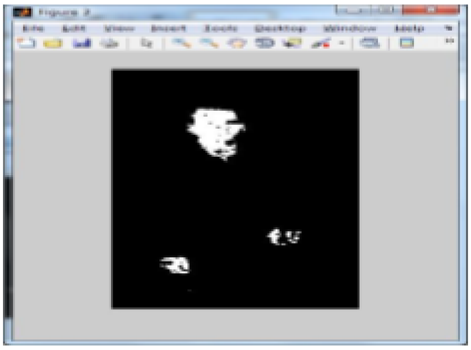

Fig.15. face detection

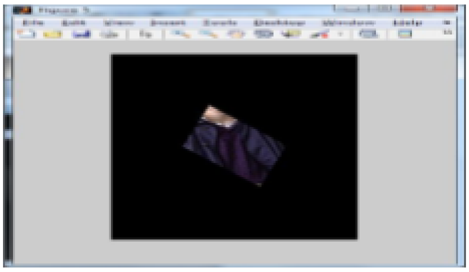

Fig.16. upper body detection

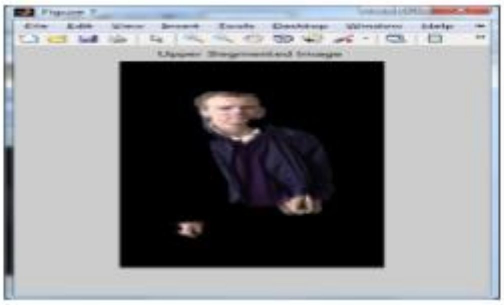

Fig.17.combination of face and upper body segmentation.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

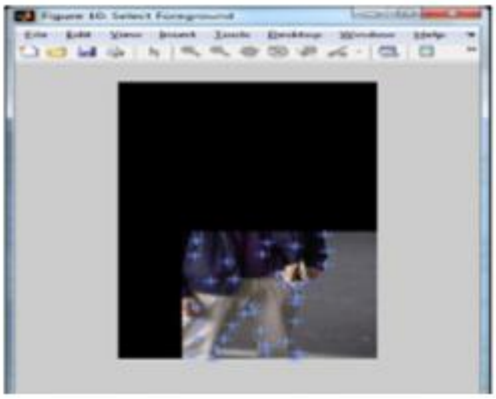

Fig.18. foreground selection

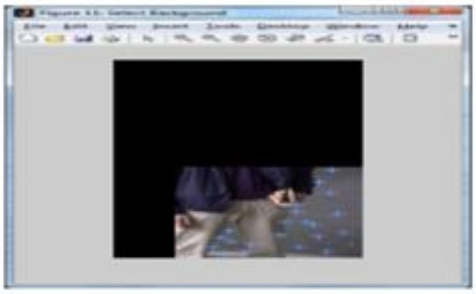

Fig.19. Background selection

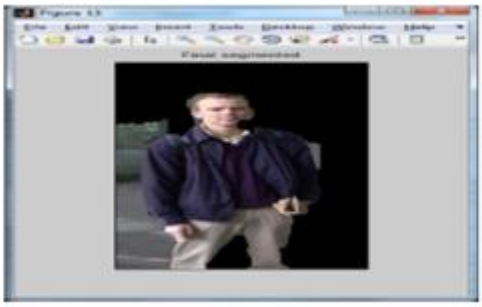

Fig.20. final output image

CONCLUSIONS

We presented a novel methodology for extracting human bodies from single images. It is a bottom-up approach that combines information from multiple levels of segmentation in order to discover salient regions with high potential of belonging to the human body. The main component of the system is the face detection step, where we estimate the rough location of the body, construct a rough anthropometric model, and model the skin's color. Soft anthropometric constraints guide an efficient search for the most visible body parts, namely the upper and lower body, avoiding the need for strong prior knowledge, such as the pose of the body. Experiments on a challenging dataset showed that the algorithm can outperform state-of-the-art segmentation algorithms, and cope with various types of standing everyday poses. However, we make some assumptions about the human pose, which restrict it from being applicable to unusual poses and when occlusions are strong. In the future, we intend to deal with more complex poses, without necessarily relying on strong pose prior. Problems like missing extreme regions, such as hair, shoes, and gloves can be solved by incorporation of more masks in the search for these parts, but caution should be taken in keeping the computational complexity from rising excessively.

REFERENCES

[1] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2009, pp. 1014–1021.

[2] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," Int. J. Comput. Vis., vol. 88, no. 2, pp. 303–338, 2010.

[3] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Progressive search space reduction for human pose estimation," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2008, pp. 1–8.

[4] M. P. Kumar, A. Zisserman, and P. H. Torr, "Efficient discriminative learning of parts-based models," in Proc. IEEE 12th Int. Conf. Comput. Vis., 2009, pp. 552–559.

[5] V. Delaitre, I. Laptev, and J. Sivic, "Recognizing human actions in still images: A study of bag-of-features and part-based representations," in Proc. IEEE Brit. Mach. Vis. Conf., 2010.

[6] A. Gupta, A. Kembhavi, and L. S. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 10, pp. 1775–1789, Oct. 2009.

[7] B. Yao and L. Fei-Fei, "Grouplet: A structured image representation for recognizing human and object interactions," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2010, pp. 9–16.

[8] P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zisserman, "Long term arm and hand tracking for continuous sign language TV broadcasts," in Proc. 19th Brit. Mach. Vis. Conf., 2008, pp. 1105–1114.

[9] A. Farhadi and D. Forsyth, "Aligning ASL for statistical translation using a discriminative word model," in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog., 2006, pp. 1471–1476.

[10] L. Zhao and L. S. Davis, "Iterative figure-ground discrimination," in Proc. 17th Int. Conf. Pattern Recog., 2004, pp. 67–70.