An Implementation of Human Body Extraction Mechanism Based on Multi-Level Image Segmentation and Spline Regression From Single Images

Y. NUTHANA (PG Scholar)[1] **A N NAGA JYOTHI** Assistant professor[2]

Department of DECS, Dr. K V Subba Reddy Institute of Technology, Dupadu, Kurnool, AP-518218, INDIA

nuthana.yerrabati@gmail.com[1] nagajyothi.adhinayakanti@gmail.com[2]

Abstract

Imaging of human body segments is demanding task which supports many applications such as understanding of scenes and recognition of activities. A bottom-up technology for extracting human bodies automatically from a single image, in case of almost upright position, is the available technique in cluttered environments. The dimension, position and face color are used for localizing human body, model construction of upper and lower body as per anthropometric constraints and skin color calculation. A highest level pose can be extracted by combining different levels of segmentation granularity. Jointly estimating the foreground and background during the body part search phase gives rise to the segments of the human body, that alleviates the need for shape matching exactly.A novel approach for extraction of standing human bodies has proposed in this paper where the highly dimensional pose space, scene density, and various human appearances are handled in a better way compared to conventional state of art methods. The proposed approach is classified into five different steps (a) face detection, (b) multi level segmentation, (c) skin detection, (d) upper body segmentation and (e) lower body segmentation respectively. Finally the simulation results have achieved better performance and high efficiency over traditional state of art methods.

Keywords: Multi level segmentation, skin detection, human bodies, super pixels, bottom-up approach

1. INTRODUCTION

Human body extraction in an unconstrained still image is a challenging task due to various factors like image noise, occlusion, and cluttered background. Knowledge about the human body can benefit various tasks, such as identification and determination of the human layout, action recognition, etc. Human body segmentation and extraction have been commonly practiced when photos and videos are available in controlled environments where we have background information. However, static images have no such uses. The problem of silhouette extraction is more challenging when the scenario is complex. Methodologies that can work at a frame level also work for frame sequences and facilitate methods of action recognition based on the body skeleton.

Another approach is to utilize some available cues to guide image segmentation to extract object. Rother et al. proposed an interactive foreground/background segmentation called GrabCut. It is an iterative image segmentation technique based upon the Graph Cut algorithm. Since the cues for image segmentation are given manually, it is mainly used as an interactive image tool for foreground object extraction.

Inspired by the work of Rother et al., we present an approach to automatically extract human body region from color photos, which incorporates dynamically updating trimap contour with iterated GrabCut technique. On considering the diversity and variety of human poses, we constrain our researches on those human poses with frontal/side faces in color photo images and focus on the topic of human body region extraction, which aims to separate human body from background and does not classify human body parts. Different from the trimap guiding the image segmentation in our approach is initialized from the results of detected faces, and the contour of the trimap is updated dynamically during body extraction. This is motivated by a fact that estimation on a small region is more accurate than on a large region if a few cues are just available. And we noticed that human torso is relatively stable in appearance compared with various human poses formed by hands and feet. A body torso is firstly extracted. Then the trimap is updated by dynamically growing its contour according to local image information, and new body region is estimated by applying GrabCut to the target image. With the iterated processing of trimap shape updating and GrabCut applying, human body region is finally extracted.

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.

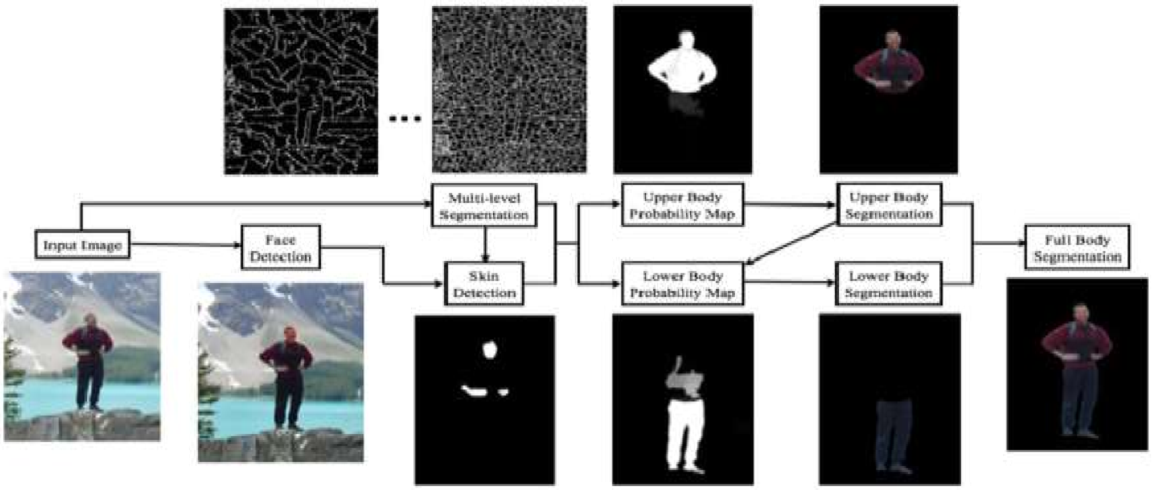

Figure 1: The proposed method methodology in different steps.

2. CONTRIBUTION

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).

3. STATE OF THE ART

The word “anthropometry’’ was coined by the French naturalist Georges Cuvier (1769–1832). It was first used by physical anthropologists in their studies of human variability among human races and for comparison of humans to other primates. Anthropometry literally means “measurement of man,’’ or “measurement of humans,’’ from the Greek words anthropos, a man, and metron, a measure. Although we can measure humans in many different ways, anthropometry focuses on the measurement of bodily features such as body shape and body composition (“static anthropometry’’), the body’s motion and strength capabilities and use of space (“dynamic anthropometry’’).

Non-rigid object detection and articulated pose estimation are two related and challenging problems in computer vision. Numerous models have been proposed over the years and often address different special cases, such as pedestrian detection or upper body pose estimation in TV footage. This paper shows that such specialization may not be necessary, and proposes a generic approach based on the pictorial structures framework. We show that the right selection of components for both appearance and spatial modeling is crucial for general applicability and overall performance of the model. The appearance of body parts is modeled using densely sampled shape context descriptors and discriminatively trained AdaBoost classifiers.

The objective of this paper is to estimate 2D human pose as a spatial configuration of body parts in TV and movie video shots. Such video material is uncontrolled and extremely challenging. We propose an approach that progressively reduces the search space for body parts, to greatly improve the chances that pose estimation will succeed. This involves two contributions: (i) a generic detector using a weak model of pose to substantially reduce the full pose search space; and (ii) employing 'grab-cut' initialized on detected regions proposed by the weak model, to further prune the search space. Moreover, we also propose (iii) an integrated spatiotemporal model covering multiple frames to refine pose estimates from individual frames, with inference using belief propagation.

4. PROPOSED METHOD

(a) Face Detection

Localization of the face region in our method is performed using OpenCV'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.

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 TGlobalSkin (set empirically to 0.7 in our experiments). If it passes the global skin test (greater than TGlobalSkin), 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.

(b) Multi 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. In this study, we propose using an image segmentation method, in order to process pixels in more meaningful groups.

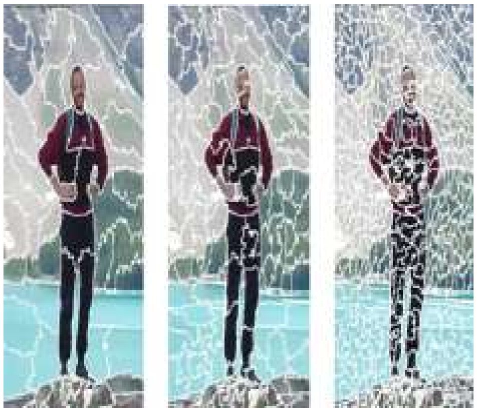

Figure 2: Image segmentation for 100, 200, and 500 super pixels

(c) Skin detection algorithm

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. 3). 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 GrabCut algorithm, which in turn outputs the final skin regions. False positives are eliminated using anthropometric constraints and body connectivity.

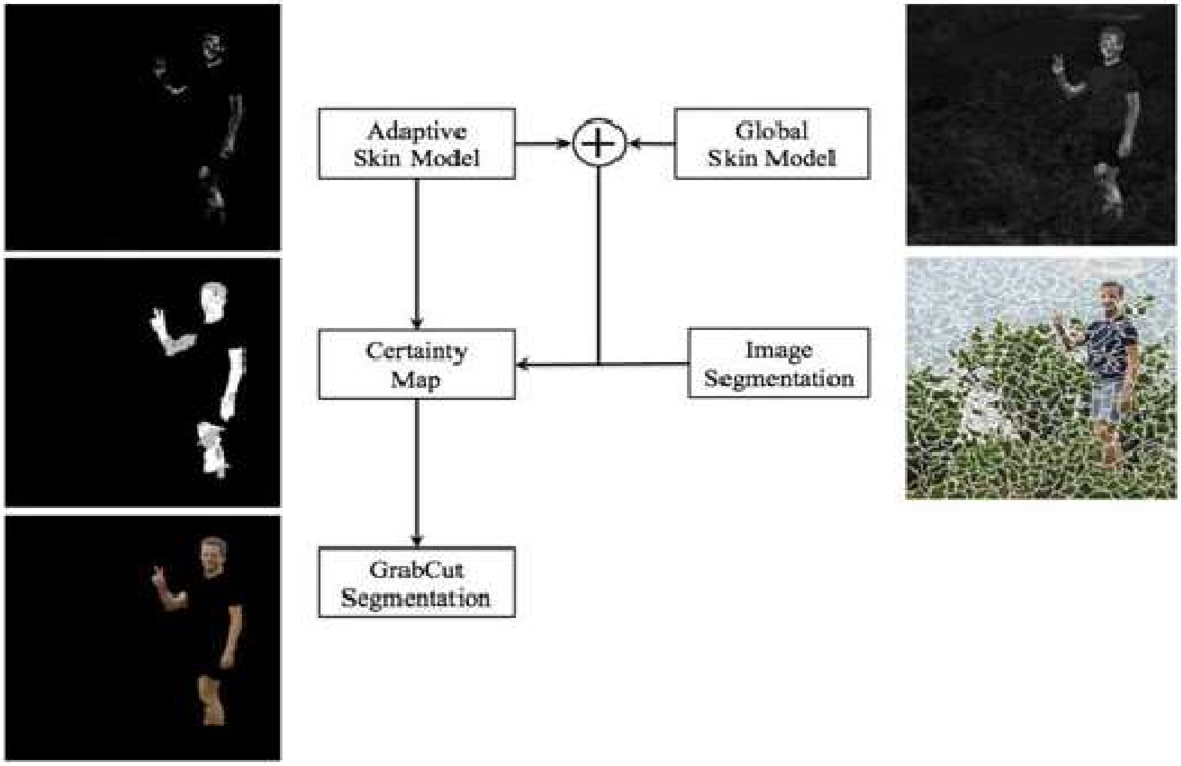

Figure 3: Skin detection algorithm

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:

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

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

(d) Upper body segmentation

In this section, we present a methodology for extraction of the whole upper human body in single images, extending , 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 superpixels instead of single pixels, which are acquired by In this section, we present a methodology for extraction of the whole upper human body in single images, extending, 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 superpixels instead of single pixels, which are acquired by an image segmentation algorithm, yield more accurate results and allow more efficient computations.

Here, we use two segmentation levels in this stage of 100 and 200 superpixels, 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.

Our approach has the advantages of taking different perceptual groupings into account and being able to alleviate the need for accurate torso mask estimation, by conjunctively measuring the foreground and background potentials. The fact that we use super pixels in the computations makes comparisons more meaningful, preserves strong boundaries, and improves algorithmic efficiency. Results may be improved by adding more segmentation levels and masks at different sizes and locations, but at the cost of computational complexity.

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. 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 (1, since the potentials are normalized). Instead of using a simple or even adaptive thresholding, we use a multiple level thresholding to recover 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 thresholdt (Region t) to the expected upper body size (ExpUpperBodySize) is calculated as follows:

$$ScoreSize = \frac{-|Region_t\_ExpUpperBodySize\ |}{ExpUpperbody} \quad (3)$$

where ExpUpperBodySize=11×PL 2 . The score for the second criterion is calculated by averaging the gradient image (GradIm) responses for the pixels that belong to the perimeter (PRegiont)of Regiont as

$$ScoreGrad = \frac{1}{|PRegion_t|} \sum^{|PRegion_t|} GradIm \cap PRegion_t \quad (4)$$

(e) 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 angleφ, translations in x- andy-axes,τx andτy and scaling inx- andy-axes,sx andsy. 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 potentialS(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 constrained. Thus, fitting of the torso rectangle is formulated as a maximization problem

$$\theta max f\,(\theta) = \alpha_1 \times UBC(\theta) + \alpha_2 \times s(\theta) + \alpha_3 \times D_{sf}\,(\theta) \quad (5)$$

where TorsoMask(θ) is the binary image, where pixels inside the rectangle rTorsoMask(θ) are 1, else 0; UBR is the binary image, where pixels inside the

International Journal of Research
Available at https://edupediapublications.org/journals

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 18
December 2016

UBR are 1, else 0; a1,a2,a3 are weights, set to 0.4, 0.5, and 0.1, respectively

5. RESULTS

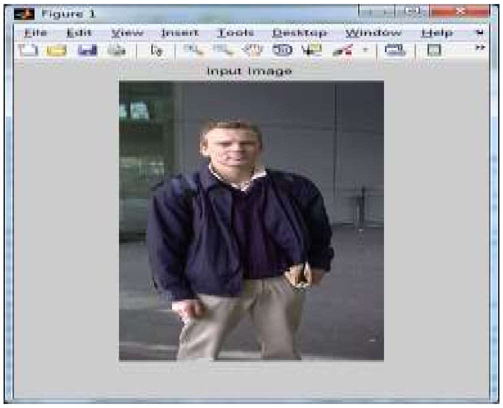

Figure 4: Input image

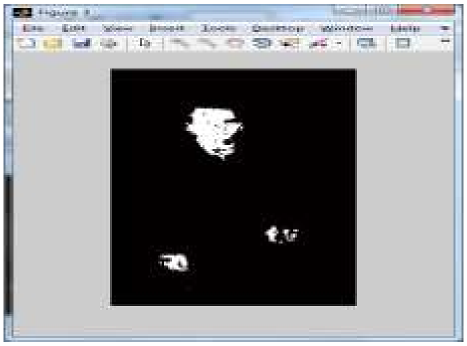

Figure 5: Face detection

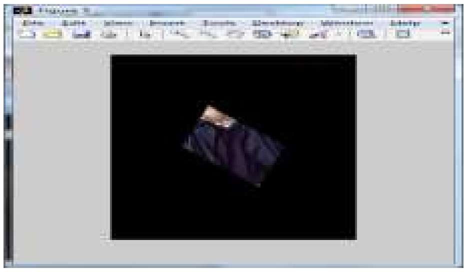

Figure 6: Rectangular method for upper body detection

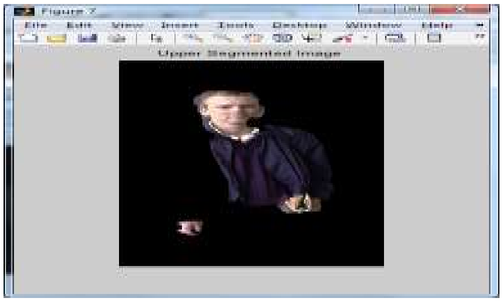

Figure 7: Collaboration of face and upper body

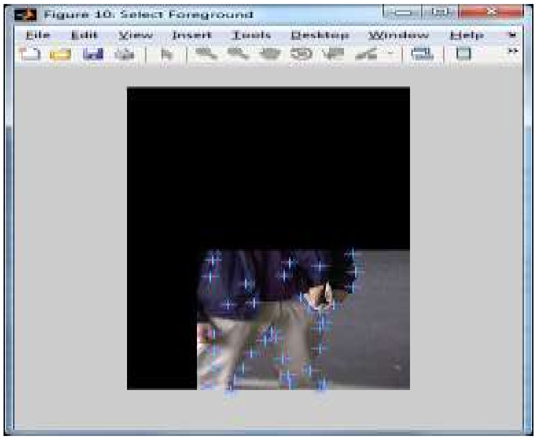
Figure 10: Select Foreground
File Edit View Insert Tools Desktop Window Help

segmentation

Figure 8: Foreground selection

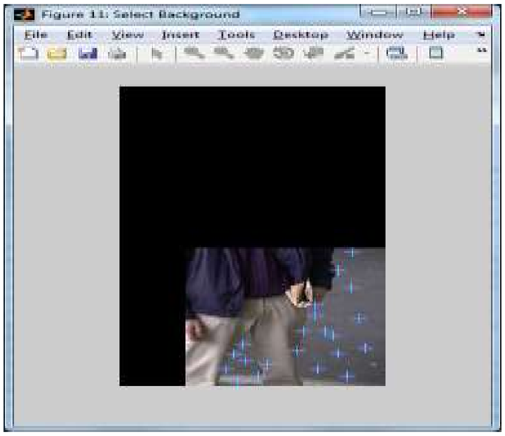
Figure 11: Select Background
File Edit View Insert Tools Desktop Window Help

Figure 9: Background selection

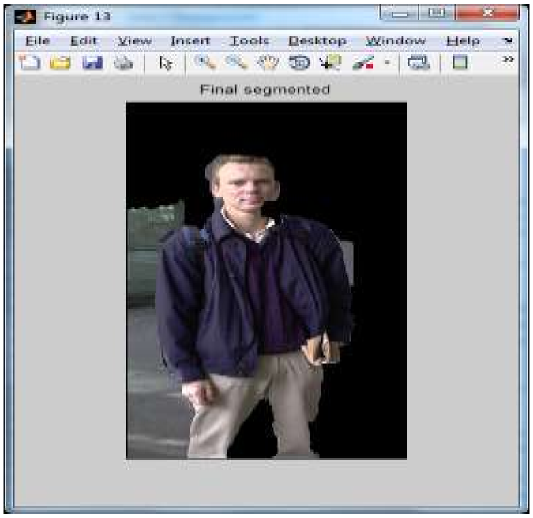

Figure 10: Final result

6. CONCLUSION

A novel approach for extraction of standing human bodies has proposed in this paper. 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.

EXTENSION

This paper is proposed to extract a standing human body by using multiple algorithms. But the time consumption is more in the proposed method. In order to avoid the computational complexity spline detector is used to improve the performance and efficiency of an image.

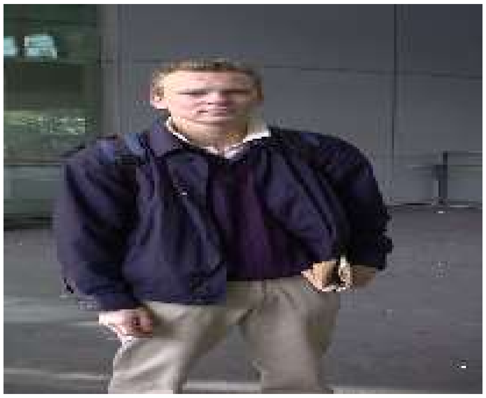

Figure 11: Input image (Spline Regression)

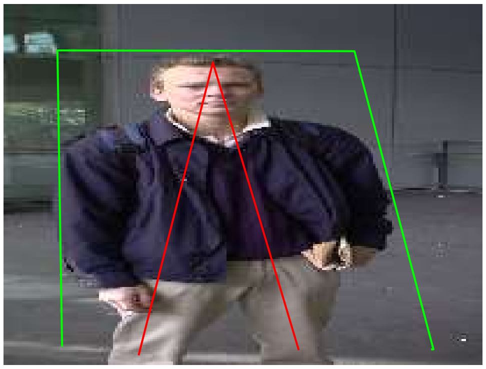

Figure 12: Placing three points on foreground (Spline Regression)

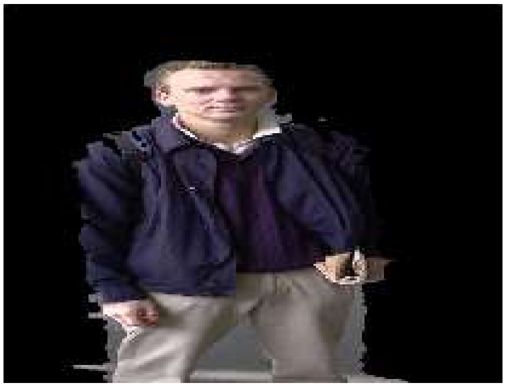

Figure 13: Final segmentation (Spline Regression)

REFERENCES

1] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," inProc. 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," inProc. 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," inProc. 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," inProc. 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," inProc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog., 2006, pp. 1471–1476.

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

[11] L. Grady, "Random walks for image segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 11, pp. 1768–1783, Nov. 2006.

[12] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," ACM Trans. Graph., vol. 23, no. 3, pp. 309–314, Aug. 2004.

[13] V. Gulshan, C. Rother, A. Criminisi, A. Blake, and A. Zisserman, "Geodesic star convexity for interactive image segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2010, pp. 3129–3136.