Where's Waldo? Human Figure Segmentation Using Saliency Maps

Omri Soceanu, Guy Berdugo, Dmitry Rudoy, Yair Moshe, Itsik Dvir

Abstract— **Human figure segmentation (HFS) is at the very core of many image and video processing tasks. Many solutions have been proposed for the separation of objects, or more specifically human figures, from image background in a video scene. Unfortunately, these solutions do not provide tight human segmentation in the most general conditions, so only a coarse segmentation of human figures can be assumed. In this paper it is assumed that a rectangular, not necessarily tight, segmentation of a human figure is available and a way to refine this segmentation is proposed. We use saliency detection for HFS in a single image, thus creating a mask that eliminates most of the background in the rectangular input while maintaining the human figure mostly intact. The proposed technique is a generalization over recently proposed saliency detection methods in order to better accommodate the special conditions of this specific problem. Tests show that saliency detection is beneficial for HFS. In addition, the proposed saliency detection is shown to improve HFS results and presents a viable solution for HFS in a single image or a video stream.**

I. Introduction

MANY image and video processing systems require accurate human figure segmentation (HFS) as a first stage, e.g., human recognition, intruder detection, video retrieval. Depending on the application, HFS should be performed on a single image or on a video sequence. Single image HFS can be performed using various methods [1]. It is hard to achieve good results using these methods due to the non-rigidity of the human body and similarly colored object and background. Video HFS is usually performed by background modeling and subtraction. Background modeling algorithms may be sensitive to shadows, lighting conditions, moving elements in the background, similarly colored object and background, etc., and therefore may fail under most general conditions [2]. Moreover, certain applications, such as image retrieval, require that the object extraction be performed from a single image, without any prior knowledge of the background or of the object. For these applications background modeling and subtraction is not applicable. Since state-of-the-art HFS techniques do not provide tight human segmentation in the most general conditions, only a coarse segmentation of human figures can be assumed.

Saliency detection presents a fast and simple method for single image object extraction. In neuroscience, an object that attracts the attention of the eye for any of many reasons is considered as salient. The result of a saliency detection system is a saliency map, which denotes a numerical value for each pixel's saliency level. Saliency can arise from contrasts between areas, for example a white dot on a black background. Sharp edges can also attract attention, and so, a circle drawn around an object in an image will result in a highly salient area. Other causes of high saliency are different orientations, intensities and so on. Saliency detection is suited to deal with HFS since it is robust under noisy conditions and only requires a single image input. It is likely to detect human figures since they receive high saliency values, as a result of the differences from their surroundings.

This paper deals with an input of a predefined rectangular bounding region, not necessarily tight, around a human figure. We assume to receive the non-accurate input from a regions-of-interest (ROI) extraction algorithm such as [3]. Saliency maps along with some assumptions about the human figure are used for producing a refined HFS inside the ROI. Figure 1 depicts possible inputs and outputs of an ideal HFS system.

The rest of this paper is organized as follows. Section II covers recent works in the field of image segmentation through saliency detection methods. The proposed HFS algorithm is described in Section III. Algorithm results are given in section IV and conclusions are drawn in Section V.

II. Recent Works

There are many approaches toward saliency detection [3, 4, 5, 6, 7, 8]. These approaches can be divided into two categories: biologically or computationally based. The biological methods attempt to reproduce the human visual stimuli system, whereas the computational methods are based on practical solutions and technology-based techniques for computation. Itti et al. [4] proposed a

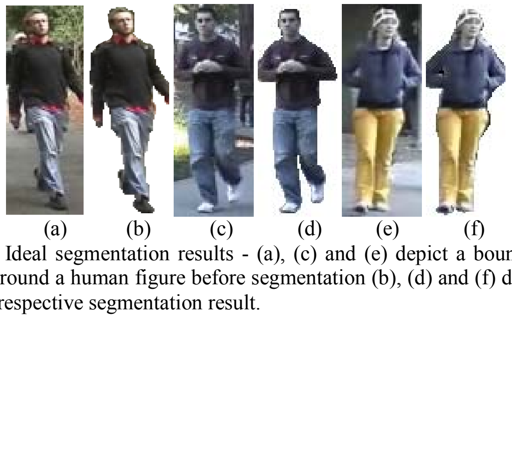

Fig.1 Ideal segmentation results - (a), (c) and (e) depict a bounding box around a human figure before segmentation (b), (d) and (f) depict their respective segmentation result.

Manuscript received November 10, 2009. This work was supported in part by Mate – Intelligent Video, 8 Hartom St. Beck Science Center, P.O. Box 45116, Har Hozvim Jerusalem 91450 Israel, www.mate.co.il.

O. Soceanu, G. Berdugo, D. Rudoy, and Y. Moshe are with the Signal and Image Processing Laboratory (SIPL), Department of Electrical Engineering, Technion - Israel Institute of Technology, Haifa, 32000 Israel, sipl.technion.ac.il (e-mails: omri123@gmail.com, guykun@gmail.com, dmitry.rudoy@gmail.com, yair@ee.technion.ac.il).

I. Dvir is with Mate – Intelligent Video (e-mail: itsik.dvir@gmail.com).

biologically based saliency detection system, which uses color, orientation and luminance at different scales in order to produce the saliency maps. However, the algorithm is computationally demanding. Ma and Zhang [5] used a computationally based algorithm that uses contrast in order to produce the saliency maps. Guo et al. [7] examined different methods for saliency map extraction, in particular, Spectral Residual Fast Fourier Transform (SR FFT) [8], Phase spectrum of Fourier Transform (PFT) and Quaternion Fourier Transform (QFT). SR FFT calculates the difference between the smoothened amplitude of FFT of the image and the amplitude itself. Then, the difference is transformed back to the spatial domain, yielding the saliency map. PFT emphasizes the role of the phase rather than the amplitude of an image in producing the saliency maps. The saliency map is extracted by transforming only the phase spectrum back to the spatial domain. Given an image pixel $I(x,y)$, its saliency level $sM(x,y)$ is computed as:

$$\phi(u,v) = \measuredangle F\left(I\left(x,y\right)\right) \tag{1}$$

$$sM\left(x,y\right) = g\left(x,y\right) * \left| F^{-1}\left[e^{j\cdot\phi(u,v)} \right] \right|^2 \tag{2}$$

Here $F(\cdot)$ and $F^{-1}(\cdot)$ denote the Fourier Transform and Inverse Fourier Transform, respectively. $\phi(u,v)$ represents the phase spectrum of the image and $g(x,y)$ is a 2D Gaussian filter. QFT improves the PFT method by using multiple frames to extract the saliency maps and therefore is not applicable in single image segmentation.

III. HFS Using Saliency Maps

In this paper we generalize the PFT algorithm in order to create an algorithm that tends to the special conditions of HFS in an ROI, i.e., a large connected-component that is more likely to be in the center of the image. Testing this generalized algorithm we demonstrate the importance of considering both the phase spectrum and the amplitude spectrum when generating saliency maps through Amplitude and Phase Spectrum of Fourier Transform (APFT) algorithm.

Extracting a human figure from a bounding box presents a unique set of characteristics. The figure is a single connected component; there is a higher probability that its comprising pixels would occupy the center of the box rather than the corners, and it most likely to be the largest object in the box. Taking these characteristics into account allows us to modify the saliency map extraction algorithm to achieve better results for HFS.

Considering the characteristics mentioned above, we modify the PFT algorithm [7] and suggest the following steps in order to produce a saliency map. Given an image pixel $I(x,y)$, its saliency level $sM(x,y)$ is computed as:

$$A\left(u,v\right) = \left| F\left(I\left(x,y\right)\right) \right| \tag{3}$$

$$sM\left(x,y\right) = f\left(x,y\right) * \left| F^{-1}\left[A^{\beta}\left(u,v\right) \cdot e^{j\cdot\phi(u,v)} \right] \right|^2 \tag{4}$$

Where $F(\cdot)$ and $F^{-1}(\cdot)$ denote the Fourier Transform and Inverse Fourier Transform, respectively. $A(u,v)$ represents the magnitude of the image and $\phi(u,v)$ represents the phase spectrum of the image as defined in Eq.1. $f(x,y)$ is a filter that represents the probability that a pixel is part of the human figure. β is a constant power by which the magnitude is raised.

Achieving good results in terms of precision and recall requires a balance between the non-parametric approach of PFT which makes no prior assumptions about the figure, and an empirical probabilistic map expressed by the filter $f(x,y)$. Such a balance is achieved by multiplying $e^{j\cdot\phi(u,v)}$ by the magnitude spectrum raised to the power of β. When $\beta = 0$ and the probabilistic map is a 2D Gaussian filter the result is identical to the result of the PFT algorithm.

Three different probabilistic filters were examined. A 2D Gaussian filter was the first obvious choice since it embodies the different characteristics mentioned previously. The highest values are present at the center of the map and gradually decrease towards the corners. The Gaussian filter is simple and its size should be set to coincide with the size of the region.

Two more refined filters based on HFS characteristics can be designed using averaged saliency maps and Multivariate Kernel Density Estimation (MKDE) [9]. In contrast to the Gaussian, the average of saliency maps requires a learning stage to take place over saliency maps produced using the PFT method. The learning stage produces a more accurate measure of the probability that a certain pixel belongs to a human figure. Pixels that receive high saliency values over numerous pictures are more likely to be part of a human figure then pixels that receive low saliency values.

MKDE can also be used to create a probabilistic model for a human figure [9]. MKDE does not make any prior assumptions, thus can converge into any probability density function. It is used in a learning stage in order to measure the probability that a certain pixel belongs to a human figure. As a kernel we take a zero mean Gaussian function with a parametrically controlled standard deviation. The following steps produce the probability map:

Given an image $I(x,y)$,

$$s_i = \left(s_{i1}, s_{i2}, s_{i3}, s_{i4}\right)^T \tag{5}$$

$$\hat{p}^{HF}\left(z\right) = \frac{1}{N_p \sigma_1 \cdot \ldots \sigma_4} \sum_{i=1}^{N_p} \prod_{j=1}^{4} \kappa\left(\frac{z_j - s_{ij}}{\sigma_j}\right) \tag{6}$$

where s_i is a 4D vector containing the saliency map extracted using the PFT algorithm of each color channel $(j = 1, 2, 3)$ and a spatial feature $(j = 4)$ for the i-th pixel. $\hat{p}^{HF}(z)$ is the probability of obtaining a given feature vector z with the same components as s_i. $\kappa(\cdot)$ denotes the Gaussian kernel, which is the kernel function used for all channels. N_p is the number of pixels in $I(x,y)$ and σ_j are parameters denoting the standard deviation of the kernels which are set according to empirical results. The aforementioned learning process is performed on a learning set of bounding boxes.

The probability density $\hat{p}^{HF}(z)$ is averaged over the learning set.

Moving from saliency maps to segmentation involves applying a threshold over the saliency maps. Pixels with saliency values greater or equal to the threshold are considered part of the human figure, whereas pixels with saliency values less than the threshold are considered part of the background. Thresholds should be set to give satisfactory results for each of the filters. Testing suggests that for the Gaussian filter the mean of the saliency intensities gives good results.

Three tests were conducted using the different probability maps. The same dataset was used for both the averaged saliency maps and the MKDE learning stages. Figure 2 shows the shape of the three proposed probability maps, where higher values are represented by higher intensity values.

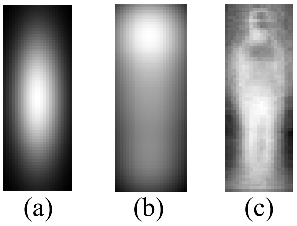

Fig. 2. Probability maps $f(x,y)$ – (a) is a Gaussian with $\sigma = 8$ (b) is derived through averaging saliency maps over a learning set of human figures bound in a box (c) is derived through MKDE averaged over the saliency maps in the learning set.

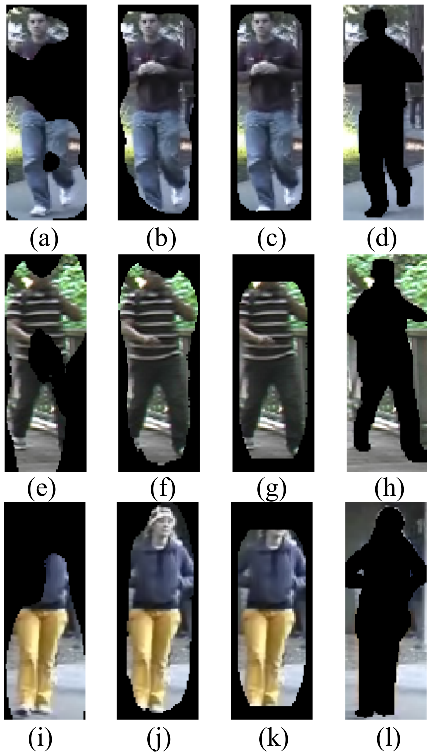

Fig. 3. HFS results using different methods. (a), (e) and (i) are results for $\beta = 0$. (b), (f) and (j) are results for β_{opt} - The β that produced the best results. (c), (g) and (k) are results for $\beta > 2$. (a), (b) and (c) are results for a Gaussian filter. (e), (f) and (g) are results produced by a filter derived through averaging saliency maps. (i), (j) and (k) are results produced using MKDE. (d), (h) and (l) are ground truth segmentations. Notice that (a) equals to results of the PFT algorithm.

IV. Results

In order to evaluate the performance of the algorithm for different parameters and for the three previously proposed probability maps, the VIPeR Dataset [10] is used. The dataset contains 632 human figures which are tightly bounded, with a fixed resolution of 48x128 pixels. For the performance evaluation, 24 frames were randomly selected from the dataset and the figures were manually segmented, as can be seen in Figure 3. For the learning stage, the rest of the VIPeR frames were used.

Different sized images produce different saliency maps. When examining a small picture one tends to overlook the small details and focus on the large objects. Therefore, before the saliency process the boxes are first scaled down to a fixed size. Thus, the human figure receives higher values than the rest of the frame in the saliency map.

The segmentation results of the different methods can be seen in Figure 3. One notices that as β increases the figure segmentation incorporates more pixels from the center of the box, and fewer pixels from the edges. This coincides with the probabilistic maps shown in Figure 2. As seen in Figure 3, the APFT method produces a better segmentation than the PFT method.

In order to evaluate the quality of the proposed algorithm numerically the following method was used [10]. Given ground-truth segmentation M and an algorithm segmentation E:

$$\begin{aligned} S_U(\beta, Th) &= \left|M - (M \cap E)\right| / |M| \times 100, \\ S_O(\beta, Th) &= \left|E - (M \cap E)\right| / |E| \times 100, \end{aligned} \tag{7}$$

$$A(\beta, Th) = 100 - (S_U + S_O) \tag{8}$$

where $|\cdot|$ denotes the number of pixels in the segmentation. S_U represents the inaccuracy of under-extraction, i.e., excluding figure-pixels from E, and S_O represents over-extraction, i.e., including background-pixels. $A(\beta, Th)$ represents the overall accuracy of the segmentation. Higher values of $A(\beta, Th)$ mean a better segmentation. We note that in boxes containing mostly figure-pixels, Eq. (8) rewards not using HFS at all, thus another quality evaluation method is used in this paper:

$$A_{Mult}(\beta, Th) = \frac{(100 - S_O) \cdot (100 - S_U)}{100} \tag{9}$$

For the Gaussian probability map standard deviation we use $\sigma = 8$. For the MKDE probability map learning stage we use the standard deviations $\sigma_1 = \sigma_2 = \sigma_3 = 4$, $\sigma_4 = 1$. Figure 4 displays precision and recall curves for the PFT algorithm and for the APFT algorithm with the three proposed probability maps. One notices that the proposed APFT algorithm outperforms the PFT algorithm, producing higher segmentation precision values for the same recall values.

Table I shows $A(\beta, Th_{opt})$ and $A_{Mult}(\beta, Th_{opt})$ values for the aforementioned probability maps. Th_{opt} is the

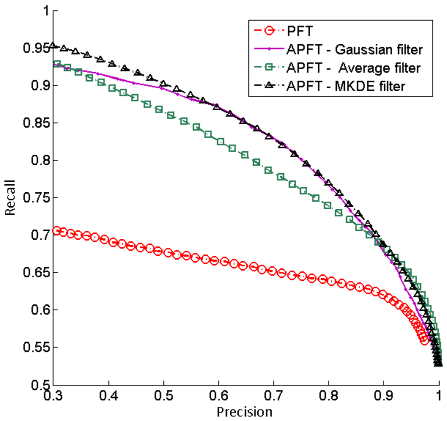

Fig.4. Precision-Recall curves for different HFS methods. For the same recall values, the precision of the segmentation is higher when using the proposed APFT algorithm compared with the PFT algorithm.

optimal threshold for each of the maps. One can notice that the MKDE filter gives a better performance than the other two probability maps. However, the Gaussian filter gives roughly the same results without any prior learning stage. The results show an improvement of up to 25% compared with the PFT algorithm. This improvement is noticeable in both the $A(\beta, Th_{opt})$ and $A_{Mult}(\beta, Th_{opt})$ results in Table I and in the precision recall curves in Figure 4.

TABLE I
HFS METHODS RESULTS

Method	$A_{Mult,opt}[\%]$	$A_{opt}[\%]$
APFT with optimal β		
Gaussian	63.48	60.00
Average saliency maps	62.50	59.84
MKDE	63.81	60.53
PFT	46.07	35.59

Examining S_U and S_O shows that at optimal $A(\beta, Th_{opt})$ values there is only a 15% under-extraction with a 25% over-extraction using the Gaussian filter with the APFT method, whereas when using the PFT method, for the same under-extraction percentile one gets a 36% over-extraction.

Examining the $A(\beta, Th)$ quality maps one notices that, as $|\beta|$ values grow, so does $A(\beta, Th_{opt})$ values, until the quality reaches its maxima, and then decreases slightly converging to a certain value. This outcome is expected due to the balancing nature of β. Testing the APFT algorithm over not tightly bounding inputs one can conclude that the β value where $A(\beta, Th_{opt})$ reaches the maxima is closely related to the tightness of the bounding boxes. In databases like VIPeR, where most of the images contain a figure in the middle of the input region, occupying more pixels than the background pixels, the optimal β value tends to be higher. Whereas when testing the APFT algorithm on inputs where the figure is not necessarily in the middle, and not occupying most of the box, the optimal β value is lower (closer to -1).

V. Conclusion

HFS is an important part in a variety of image processing systems. Given a bounding box around a human figure, several methods were examined trying to improve segmentation in a single image. Taking into consideration the unique properties of HFS we have managed to improve segmentation results. Previous works were generalized and produced good results. It is proposed to use different filters in the saliency production process as well as to incorporate the image magnitude into the process. Best results were achieved with a filter derived using MKDE on saliency maps. However only slightly less accurate results were obtained using a simple Gaussian filter which requires no prior learning stage or complex computations. Multiplying by the magnitude spectrum in order to balance between the probabilistic model and the saliency map was shown to be a decisive factor in the improvement of the HFS.

Acknowledgment

We would like to thank Eliran Subag for the mathematical advice. We would also like to thank the staff of the Signal & Image Processing Lab (SIPL) for their support, and especially its head, Prof. David Malah.

References

[1] T.B. Moeslund, A. Hilton, and V. Krüger, “A survey of advances in vision-based human motion capture and analysis”. *Computer Vision and Image Understanding*. 104, 2, 90–126 (2006).

[2] S. Elhabian, K. El-Sayed and S. Ahmed, “Moving object detection in spatial domain using background removal techniques”. Recent Patents on Computer Science 1, 32–54 (2008).

[3] B. C. Ko and J.-Y. Nam., “Object-of-interest image segmentation based on human attention and semantic region clustering”. *Journal of Optical Society of America* A, 23(10):2462-2470, October 2006.

[4] L. Itti, C. Koch and E. Niebur: “A model of saliency-based visual attention for rapid scene analysis”. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11), 1254–1259 (1998)

[5] Ma, Y.-F., Zhang, H.-J.: “Contrast-based image attention analysis by using fuzzy growing”. In: Proceedings of the Eleventh ACM International Conference on Multimedia, November 2003, pp. 374–381 (2003)

[6] R. Achanta, F. Estrada, P. Wils and S. Süsstrunk, ”Salient region detection and segmentation”, in: A. Gasteratos, M. Vincze, J. K. Tsotsos (Eds.), Computer Vision Systems, Vol. 5008 of LNCS, Springer, 2008, pp. 66–75.

[7] C. Guo, Q. Ma and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform" presented at *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), pp 1–8, 2008.

[8] X. Hou and L. Zhang, “Saliency detection: A spectral residual approach”, in *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR07). IEEE Computer Society, June 2007.

[9] D. W Scott, “*Multivariate Density Estimation: Theory, Practice, and Visualization.”* New York: Wiley (1992).

[10] D. Gray, S. Brennan and H. Tao, "Evaluating Appearance Models for Recognition, Reacquisition, and Tracking," Performance Evaluation of Tracking and Surveillance (PETS). IEEE International Workshop on, 2007.