Enhancing Video Object Segmentation Results Through Biologically Inspired Postprocessing

Dubravko Ćulibrk+, Vladimir Radenković*, Daniel Socek†

***Abstract*— Object segmentation from a video stream is an essential task in video processing and forms the foundation of scene understanding, object-based video encoding (e.g. MPEG4), and various surveillance and 2D-to-pseudo-3D conversion applications. Many segmentation approaches are pixel-based and suffer from noise in the segmentation results, due to the fact that these approaches do not exploit spatial information. Morphological post processing is typically used to enhance the segmentation results. Here, an alternative post-processing approach is presented. The proposed approach is inspired by well-known aspects of the primate visual system function. The approach is particularly suitable for use with the probabilistic segmentation algorithms and allows for efficient neural-network-based implementation.**

***Index Terms*— Object segmentation, Post processing, Video processing, Background subtraction, Automated surveillance.**

I. Introduction

OBJECT segmentation is an essential task in the domain of digital video processing. Diverse applications, such as scene understanding, object-based video encoding, surveillance applications and 2D-to-pseudo-3D video conversion, rely on the ability to extract objects from video sequences.

The research into object segmentation for video sequences grabbed from a stationary camera has yielded a number of approaches based on the detection of the motion of objects. The approaches of this class scrutinize the changes observed between the consecutive frames of the sequence to detect pixels which correspond to moving objects. The task is particularly difficult when the segmentation is to be done for natural scenes where the background contains shadows, moving objects, and undergoes illumination changes.

For purposes of automated surveillance and scene understanding, it is often of more interest not only to detect the objects moving in the scene, but to distinguish between two classes of objects:

- *Background objects* corresponding to all objects that are present in the scene during the whole sequence or longer than a predefined period of time.
- *Foreground objects* representing all other objects appearing in the scene.

The goal of the foreground segmentation is to separate pixels corresponding to foreground from those corresponding to background.

The most general and widely used approaches to foreground segmentation are based directly on the definitions of segmentation entities as stated above. To achieve segmentation they model one of the entities (usually background) in time and classify pixels in each frame of the sequence through comparisons with the expected traits of the pixels of modelled entity.

If the state of the background is known for every frame of the sequence and there are no changes in illumination, the segmentation can be accomplished by a simple comparison between the background image and a frame of the sequence. This, however, is unrealistic for almost all applications. In the absence of an exact model for the background, one has to be estimated based on the information in the sequence and some assumptions. The process of modeling the background and determining the foreground by comparison with the frames of the sequence is often referred to as *background subtraction*.

Background models used differ from one approach to the other. With respect to the granularity of the models, two broad classes of background-subtraction methods can be identified:

1) Region based models.
2) Single-pixel-based models.

The single-pixel-based models treat the sequence as a set of single-pixel values changing over time. Stauffer *et al.* [1] refer to this paradigm as a pixel process. Within this class of background-subtraction methods, the statistics of pixel processes are gathered for each process separately and then used to classify each pixel independently.

While single-pixel-based approaches achieve best segmentation results, they suffer from the inherent deficiency of relying solely on temporal information extracted from the sequence. This results in segmentation results that are cluttered with small objects falsely segmented as foreground and foreground objects that are not segmented as solid, due to foreground pixels falsely labelled as background. The effect is especially pronounced when natural scenes containing complex illumination changes and motion in the background (eg. due to waves, trees moving in the wind, etc.).

To alleviate the effects of such segmentation errors most approaches incorporate some sort of segmentation result postprocessing. This is done either through morphological operations or modelling of the errors produced by the initial segmentation. Once the segmentation errors are removed, the information about the false classified pixels can be fed back to the initial segmentation algorithm to enhance the initial segmentation in the future.

In the work presented here, a novel approach to enhancing segmentation results is described. It is based on the fact that the objects in the scene are expected to occupy larger number of pixels and that the concentration of pixels classified as foreground is greater in the regions occupied by foreground

+dculibrk@fau.edu, *vrftn@uns.ns.ac.yu, †dsocek@fau.edu
Authors are with University of Novi Sad, Serbia and Florida Atlantic University, USA

objects, regardless of the error produced by the initial segmentation. The approach is inspired by receptive fields of neurons in the visual cortex of primate brain and uses a negative Mexican hat filter to enhance the segmentation results. This allows for efficient exploitation of the spatial information contained in the frames of the sequence, as well as efficient implementation.

The final segmentation results achieved have been compared to those obtained through complex morphological filtering. Preliminary findings, based on a diverse set of surveillance related sequences compiled by Li *et al.* [2] are presented in this paper. The approach presented has been proven to achieve better final segmentation results, be more robust to the variation of parameters and is of lower computational complexity when compared to traditional morphological filtering.

As primary segmentation method in the work presented here, Background Modeling Neural Network approach to foreground segmentation [3] was used. The approach is a probabilistic, single-pixel-based background-subtraction method that relies on neural networks to model the background and perform the segmentation.

Section II provides a survey of related published work. Section III describes the particulars of the proposed approach. Section IV is dedicated to the presentation and discussion of experimental results. Section V contains the conclusions.

II. Related Work

A general discussion of the state of the art background-subtraction approaches can be found in [3]. The approach presented here is primarily intended to enhance the segmentation performed by the single-pixel-based background-subtraction algorithms. [4] holds several references to region-based and frame-based segmentation approaches.

Following the classification in [3], the discussion of the segmentation approaches and postprocessing steps taken to enhance the segmentation results, is limited to probabilistic approaches, since the simpler, filter-based approaches, are not suitable for use in complex-background scenarios where large amounts of noise are to be expected [1] [3] [2] [5].

A well-known approach to background modelling and subtraction are Mixture of Gaussians (MoG) models. These methods use multiple evolving Gaussian distributions as a model for the values of the background pixels [6] [1] [7]. While Stauffer *et al.* [1] achieved adequate final segmentation through the use of connected components algorithm, that allowed for the removal of small components, Li *et al.* had to resort to more elaborate segmentation-result-enhancement framework to correct the results of the segmentation achieved by MoG on their test set [2].

A nonparametric kernel density estimation framework for foreground segmentation and object tracking for visual surveillance has been proposed in [5]. The authors used an elaborate second stage probabilistic algorithm to model the probability of a pixel labelled as foreground, being the result of a background object moving further away than previously observed. They proposed the use of morphological filtering to remove any remaining noise.

Li et al. proposed an alternate single-pixel-based segmentation approach [2]. Initial segmentation is achieved through comparison with a reference background image stored as a model of the background. The authors model the PDFs of the pixel values detected by the initial segmentation. Thus, they are able to distinguish between certain types of errors in the initial segmentation and the true foreground pixels. The PDFs are updated in time and a Bayes' rule based decision framework is used for classification. However, their probabilistic framework modelled the errors at single-pixel level, resulting in noisy segmentation. The authors used a sequence of morphological opening and closing operations, as well as comparison between adjacent frames to enhance the segmentation performance of their approach. They made such postprocessing an inherent part of their algorithm, by using the final result to update the learnt PDFs.

More recently (2005), Socek *et al.* [8] created a hybrid algorithm for video object segmentation in maritime surveillance domain. The approach utilized a color-based still image Felzenschwalb-Hüttenlocher (F-H) segmentation algorithm to enhance the change-based segmentation achieved through the use of a simplified algorithm of Li *et al.*. The F-H algorithm was tuned to over-segment each frame of the original sequence, creating large number of small blobs. Simple filtering was then performed by assigning each blob to foreground (background) if it contained a larger portion of foreground (background) pixels. Better results have been achieved on a limited set of test sequences, since the F-H algorithm makes better use of spatial information contained in the frame, than the morphological framework used by Li *et al.*. However the use of F-H algorithm made the approach computationally expensive.

Cristiani *et al.* [4] propose an enhancement to the MoG scheme, that is aimed at exploiting the spatial information for the enhancement of the segmentation. Their proposed scheme assumes a Gaussian shaped influence between the neighboring pixels, but takes into account only the pixels that are similar in color (greyscale) values to the pixel being processed. They report improved segmentation results for the Wallflower [9] test data set.

The approach used for evaluating the effects of segmentation enhancement filtering proposed in this paper is that of Background Modelling Neural Networks (BNNs) as described in [3]. The networks represent a biologically plausible implementation of Bayesian classifiers and nonparametric kernel based density estimators. The weights of the network store a model of background, which is continuously updated. The PDF estimates consist of a fixed number of kernels, which have fixed width. The appropriate width of the kernels was initially determined empirically, but the BNNs have recently been extended to incorporate adaptive kernel width estimation [10]. Results superior to those of Li et al. and MoG with 30 Gaussians are reported in [3]. The comparison is based on the use of morphological filtering similar to that used by Li *et al.*. The authors have made the Matlab scripts used for segmentation results postprocessing publicly available at `http://mlab.fau.edu`.

The segmentation results enhancement approach presented

here is inspired by the receptive properties observed for certain neurons of a primate visual system. For a good overview of the function of primate visual system and the neural network based simulation of the visual processing achieved by the same system the reader is referred to [11].

III. Spatial Filtering of Segmentation Results

The goal of the postprocessing step of segmentation algorithms is to reclassify pixels falsely detected as background (foreground) based on the use of additional information, not taken into account by the initial segmentation. The additional information is due to the assumption that the objects, or parts of their surfaces observed in the scene are somewhat homogenous and span the region larger than a single pixel. Clearly, this assumption can easily be made for most applications.

The filtering performed by the proposed approach is inspired by the fact that the receptive fields of a number of neural cells within primate visual system are such that the cell is excited by stimuli presented on the border of its receptive field, while same stimuli in the center of the field inhibit the response. It is proposed here that one of the effects of such architecture is that the system is robust to high levels of noise occurring in otherwise homogenous regions. By creating the filter that mimics the cell sensitivity over the receptive fields, one is able to perform morphological filtering that can be effectively applied to enhance segmentation results in the domain of foreground detection.

A negative Mexican hat function can be applied to describe the effects of the stimuli presented to the cell [11], over the receptive region of the cell:

$$f(x) = -\frac{2}{\sqrt{3}}\pi^{-\frac{1}{4}} \cdot (1-x) \cdot \exp\frac{-x}{2} \qquad (1)$$

where x represents the square of the Euclidean distance of the point from the center of the receptive field. Figure 1 shows the plot of such a function over a square neighborhood 11 pixels wide.

The proposed algorithm for the enhancement of segmentation results operates on binary images in which the foreground is indicated with white pixels (value 1), while the background is black (value 0). Such a frame is shown in Figure 2(a). The algorithm is as follows:

1) Convolute the image with the filter of size α.
2) Normalize the filtered image to the $[0, 1]$ range.
3) Convolute the normalized image with the filter based on Equation 1 of size α.
4) Normalize the filtered image to the $[0, 1]$ range and discard all values lower than the threshold Θ.

The algorithm has two parameters, corresponding to the size of the filter α used and the threshold Θ used to determine which part of the resulting filtered image will be considered foreground. The same filter is applied twice to sufficiently attenuate the noise in the segmented image. Figure 2 shows the initial (not processed) segmentation result (Figure 2(a)), intermediate results after each stage of filtering (Figures 2(b) and 2(c)) and the final segmentation result (Figure 2(d)) obtained for $\alpha = 11$ and $\Theta = 0.35$. Figure 2(e) indicates the segmentation of the frame achieved through morphological postprocessing previously used in [3].

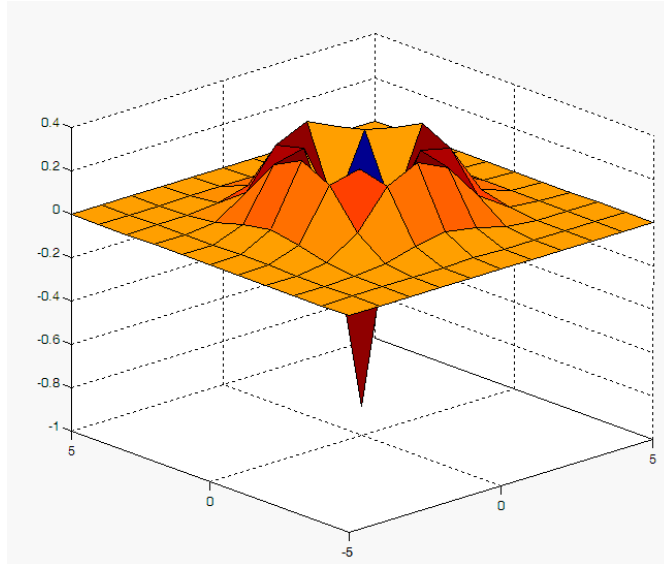


Fig. 1. Plot of the receptive field model function.

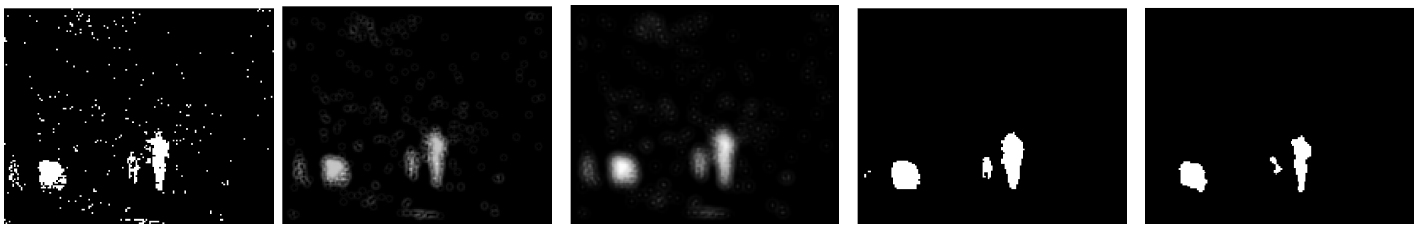
IV. Experiments and Results

To test the approach proposed we used segmentation results publicly available at `http:\\mlab.fau.edu` as noted in [3]. These frames provided were obtained as a result of segmentation of a set of diverse sequences containing complex background conditions, provided by Li *et al.* [2] and available at `http://perception.i2r.a-star.edu.sg`. The ten testing sequences were obtained in several different environments. They can be classified based on the sources of complexity in background variation, pertinent to each environment. Three classes of sequences (environments) can be identified: outdoor environments, small indoor environments and large (public) indoor environments. For each of the ten sequences Li *et al.* provided ground-truth frames, which were obtained by segmenting the foreground objects manually.

The segmentation results postprocessing performed by Culibrk *et al.* [3] consisted of an elaborate framework of morphological open and close operations, with the square support region 3 pixels wide, a connected components algorithm and removal of small regions and a hole filling operation. The postprocessing was tuned to each separate sequence, since the segmentation algorithm tended to produce different levels of noise for different sequences. The postprocessing and comparison Matlab scripts used have been made available at `mlab.fau.edu`.

The frames shown in Figure 2 are pertinent to a video of a campus driveway. The complexity of the background in this sequence is due to the trees in the background moving violently in the wind and due to the changing illumination.

To express the quality of the final segmentation result quantitatively, we follow the approach of Li *et al.*. A similarity measure of the segmentation accuracy is calculated following the methodology used in [12]. If D is a detected(segmented)



(a) Initial segmentation result. (b) After first-stage filtering. (c) After second-stage filtering. (d) Final segmentation result. (e) Li *et al.* result.

Fig. 2. Final filtering using morphological operations.

TABLE I
SIMILARITY MEASURE VALUES FOR EACH TEST SEQUENCE.

	CAM	SW	FT	WS	MR
Morphology	0.5256	0.6216	0.4636	0.7540	0.7368
Proposed approach	0.5462	0.6446	0.4225	0.7746	0.7653
	LB	SC	AP	BR	SS
Morphology	0.6276	0.5696	0.3923	0.4779	0.4928
Proposed approach	0.4225	0.5667	0.4554	0.4459	0.6231

region and G the corresponding ground truth, then the similarity measure between these two regions is defined as:

$$S = \frac{D \cap G}{D \cup G} \tag{2}$$

The similarity of the regions (S) reaches the maximum value of 1 if they are the same. Otherwise, it varies between 1 and 0 according to how similar the regions are. S is the measure of the overall misclassification.

Table I shows the similarity measure values obtained for each sequence for the proposed approach and the values previously published in [3]. The overall similarity obtained for the proposed approach (0.568) is slightly higher than that obtained for the original morphological operation framework (0.566). The size of the proposed filter was the same for all sequences (6 pixels) and the threshold value was set to a 0.5 for four of the sequences (WS, MR, AP, FT) and 0.45 for other 6 sequences. No special tuning was attempted for each specific sequence. The values for the two parameters were selected empirically. The performance proved stable over a range of filter sizes and the enhanced/deteriorated monotonically when moving away from the optimal threshold value.

V. CONCLUSION

A novel approach to the enhancement of foreground segmentation results has been presented. The approach is biologically inspired and represents an efficient way to utilize spatial information contained in the video sequence to enhance the performance single-pixel-based background subtraction methods. A negative Mexican hat filter has been used to increase the homogeneity of the segmented regions and therefore remove noise due to single-pixel-based background subtraction.

The approach was evaluated on frames segmented for a set of diverse sequences, pertinent to the automatic surveillance application domain. Good final segmentation results have been obtained for these complex sequences. The proposed approach represents an improvement in segmentation ability when compared to traditional morphological processing and required significantly less tuning to the properties of a specific sequence. In addition, the computational complexity of the postprocessing has been significantly reduced due to the elimination of complex algorithms such as hole filling and connected components determination.

The preliminary results presented in the paper indicate that the performance is robust with respect in to the values of the two parameters (α and Θ). Further studies should be conducted to se whether the approach would benefit from adaptive threshold estimation. In addition, the performance of the approach should be verified for segmentation methods other than background modelling neural networks.

REFERENCES

[1] C. Stauffer and W. Grimson, "Learning patterns of activity using real-time tracking," in *IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, pp. 747-757*, 2000.

[2] L. Li, W. Huang, I. Gu, and Q. Tian, "Statistical modeling of complex backgrounds for foreground object detection," in *IEEE Trans. Image Processing, vol. 13, pp. 1459-1472*, 2004.

[3] D. Culibrk, O. Marques, D. Socek, H. Kalva, and B. Furht, "Neural network approach to background modeling for video object segmentation," in *IEEE Trans. on Neural Networks, vol. 17, in press*, 2007.

[4] M. Cristani and V. Murino, "A spatial sampling mechanism for effective background subtraction," in *Proc. VISAPP 2007, vol. 2, pp. 403-410*, 2007.

[5] A. ElGammal, R. Duraiswami, D. Harwood, and L. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," in *Proc. of the IEEE, vol. 90, No. 7, pp. 1151-1163*, 2002.

[6] T. Ellis and M. Xu, "Object detection and tracking in an open and dynamic world," in *Proc. of the Second IEEE International Workshop on Performance Evaluation on Tracking and Surveillance (PETS'01)*, 2001.

[7] L. Ya, A. Haizhou, and X. Guangyou, "Moving object detection and tracking based on background subtraction," in *Proc. of SPIE Object Detection, Classification, and Tracking Technologies, pp. 62-66*, 2001.

[8] D. Socek, D. Culibrk, O. Marques, H. Kalva, and B. Furht, "A hybrid color-based foreground object detection method for automated marine surveillance," in *Proc. of the Advanced Concepts for Intelligent Vision Systems Conference (ACIVS 2005)*, 2005.

[9] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," in *In Proceedings of IEEE Int'l Conf. On Computer Vision, pp. 255-261*, 1999.

[10] D. Culibrk, O. Marques, D. Socek, and B. Furht, "Automatic kernel width selection for neural network based video object segmentation," in *Proc. VISAPP 2007, vol. 2, pp. 472-479*, 2007.

[11] E. M. Rolls and G. Deco, *Computational Neuroscience of Vision*. Oxford Univeristy Press, 2002.

[12] L. Li, W. Huang, I. Gu, and Q. Tian, "Foreground object detection from videos containing complex background," in *Proc. of the Eleventh ACM International Conference on Multimedia (MULTIMEDIA'03), pp. 2-10*, 2003.