Graph-Based Superpixel Labeling for Enhancement of Online Video Segmentation

Alaa E. Abdel-Hakim

Electrical Engineering Department
Assiut University
Assiut, Egypt
alaa.aly@eng.au.edu.eg

Mostafa Izz

Cairo University
Cairo, Egypt

mostafaizz@eng.cu.edu.eg

Motaz El-Saban

Microsoft Research
Cairo Advanced Technology Lab
Cairo, Egypt
motazel@microsoft.com

***Abstract*— In this paper, we propose a novel approach for video segmentation. The proposed work is based on exploiting a superpixel-based image segmentation approach to improve the performance of state-of-the-art foreground/background segmentation techniques. A fusion between a bilayer segmentation and a geodesic segmentation approaches with a graph-based superpixel segmentation method is performed. Four different combination alternatives are investigated in terms of performance and efficiency. Manually-labeled ground truth video sequences as well as our own recorded video sequences were used for evaluation purposes. The evaluation results confirm the potential of the proposed method in enhancing the accuracy of the video segmentation over the state-of-the-art.**

Keywords*— *Video Segmentation, Background Seperation, Superpixel, Bilayer, Geodesic

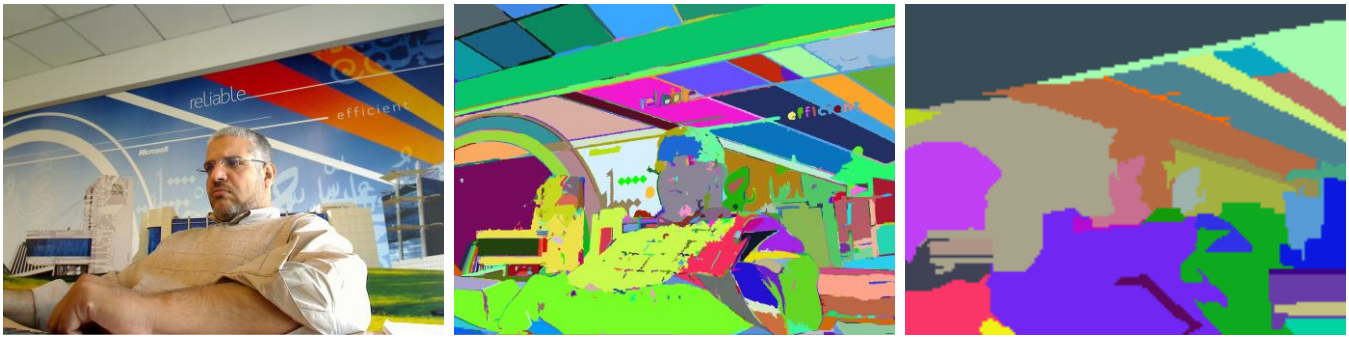
I. INTRODUCTION

Online video segmentation has gained a lot of interest during the past few years due to its importance in many applications. However, the accuracy of segmentation is still questionable. Some attempts exist in literature to improve the segmentation accuracy. For example, Kolmogorov et.al. [7] have made use of the extra information that is provided by stereo through a binocular video segmentation method. Almost all preceding research studies dealt with monocular video segmentation.

Background, appearance, motion, and image contrast have been proposed as clues for video segmentation. Different parameters are exploited in an energy equation, which can be solved efficiently with min-cut []. Energy minimization techniques have been used in [8, 13, 12,] but with different approaches. In [], a motion model for the foreground object with color model was trained to create probabilistic model for the foreground objects. In [13, 12], the concept of motons has been added with tree-based classification model to get results comparable to the binocular segmentation. Lien and Wang [8] proposed a model that combines motion, shape and color in an energy minimization function but with almost no prior knowledge with the assumption of single-type objects in the foreground.

Research studies have been presented to deal with background modeling. Monnet et.al.[9] proposed a background modeling method to work with activities in the background. Sun et.al.[10] presented some strategies to deal with sudden variations caused by illumination changes and camera shaking. Yin et.al.[12] developed an algorithm that uses trained classifier to segment moving background objects from foreground ones. In [5], a method was proposed to handle camera rotation in the video segmentation problem. The proposed approach requires a panoramic image background preconstruction for background subtraction and online segmentation. This constraint is hard to be fulfilled for the moving camera case because of the difficulty of both of panorama construction and registration in this case.

The segmentation of a certain frame can be very useful for the immediate consecutive frame. This is because of the continuity assumption under which adjacent frames can be assumed to be very similar. Some segmentation methods made use of this fact, like the interactive video segmentation methods. In such methods, the continuity assumption is exploited by optimization over the 3D video cube [11]. Nonetheless, such methods are infeasible for online segmentation. Criminsi et.ala. [] have proposed a technique like the one in [11] but using geodesic-based segmentation instead of a min-cut approach, which greatly improved segmentation speed. Nonetheless, it still cannot be applied to live videos because of the availability requirement of the entire sequence frames for segmentation. The work of Zhong et.al. [14] was based only on the continuity assumption by propagating segmentation from one frame to another. It combines temporal prior and local color distribution with a geodesic-based approach for the final segmentation. This yields a method whose efficiency is much better than min-cut. Yin et.al.[13] extended their earlier work of [12] and claimed that all previous models were not as efficient as the stereo-based video segmentation. So, they introduced conditional random field model that makes use of a fusion of shape, motion, color, and contrast with local smoothness prior. This fusion achieves segmentation through min-cut that is comparable to the stereo-based segmentation and that requires no initialization.



(a) **(b)** **(c)**

Figure 1: Superpixel operation after applying both of superpixel clustering algorithm [6] and the proposed merger algorithm. (a) The input frame (b) Clusters generated by the algorithm of [6] (c) Clusters generated by the proposed algorithm.

In this paper, the contribution lies in presenting a novel framework for accurate video segmentation. The proposed framework does not need extra hardware, like stereo-based models, nor pre-training or shape constraints. The proposed approach makes use of superpixel-based segmentation methods to enhance the accuracy of the video segmentation. The proposed approach fuses the segmentation solution of the bilayer segmentation approach [13], Geos [], and the superpixel segmenter [6]. The qualitative and quantitative evaluation results show how this kind of fusion improves the segmentation accuracy of the state-of-the-art.

II. VIDEO SEGMENTATION

In this section, we explain the video segmentation methods on which we build our proposed approach.

A. Bilayer Segmentation

Bilayer segmentation [12] uses temporal information to calculate the probability of a pixel being foreground or background. Given previous frames and previous frames foreground/background segmentation mask, in monocular video sequence, video segmentation can be performed. The final model of bilayer segmentation fuses motion color and contrast cues probabilistically with temporal and spatial priors to get accurate foreground/background segmentation. For more details of bilayer segmentation, the reader is referred to [12].

B. Geodesic Image Segmentation

Geodesic Image Segmentation (GeoS) [] uses geodesic distance transform in the gradient image to determine the related pixels and group them in a segment. Using the morphological geodesic operators, a segmentation method that is comparable to graph-cuts is obtained. The segmentation takes overhead of few milliseconds to calculate the boundaries of different segments in the images. For more details of bilayer segmentation, the reader is referred to [].

C. Superpixel-Based Segmentation

Superpixel-based segmentation depends on clustering "similar" pixels into clusters called "superpixels". Various methods of superpixel-based segmentation were presented in the literature, e.g. [6]. The work of Felzenszwalb and Huttenlocher [6] is an example of graph-based superpixel inference. As a graph-based image segmentation technique, the problem is represented in terms of an undirected graph $G=(V;E)$ where image pixels are represented by nodes (V), from which groups are connected in pairs by edges (E). In [6], edges connect pairs of neighboring vertices. Weights $w(v_i;v_j)$ are assigned to each edge $(v_i;v_j)$. These weights represent a non-negative measure of the dissimilarity between nodes v_i and v_j. The dissimilarity between nodes is measured by the difference in local attributes like intensity, motion, color, location or any other attribute.

Superpixel clustering is then done by segmenting the graph G. Figure 1 shows an example of superpixel segmentation using the graph-based approach of [6]. For the detailed algorithm, the reader is recommended to return to [6].

For the problem under consideration: video segmentation, the graph-based approach of [6] has a critical shortcoming. That is the generated graphs usually contain fine details, which are undesirable in our case. Specifically, as bilayer is a pixel-based approach, an unsupervised clustering of more pixels in larger superpixels is expected to add to the original bilayer approach. Otherwise, smaller superpixels tend to be like the original pixel-based bilayer; hence minimal impact is obtained from the fusion process. Therefore, any unsupervised clustering algorithm can contribute to this point.

So, we developed a merger algorithm to produce coarser superpixels. The developed algorithm depends on iterative reevaluation of the nodes similarity measure. The similarity is evaluated for the mean colors of the neighboring superpixels. If two nodes are "similar enough," the two nodes are merged into a larger superpixel. Figure 1-c shows an example

applying the merger algorithm on the superpixels that were obtained in Fig. 1-b. It is noticed that the number of the foreground pixels is decreased. This allows better fusion results. Nonetheless, over-merging may result in worse segmentation, especially in existence of similar neighboring foreground and background regions, as illustrated in background building picture behind the left shoulder of Fig. 1. Therefore, a tradeoff is needed depending on the nature of the imaging circumstances, in terms of the background/foreground relation, of the input sequences.

III. BILAYER-GEOS SEGMENTATION

The proposed bilayer/Geos segmentation approach enhances the segmentation performance by utilizing the pros of each of the core techniques. Specifically, for bilayer it is accurate in solving foreground/background segmentation in video sequences in the inner areas. However, it suffers from noticeable errors at the boundaries beside its low efficiency that is represented in large processing time.

On the other side, the Geos segmentation is accurate in segmenting single image into foreground and background continuous pixels with smooth boundaries. Nonetheless, it needs an initial mask to follow.

We use bilayer as a first step in our segmentation method after downscaling the input frames. This step produces an input mask to Geos. This mask is upsampled again before using Geos. Then, this mask is used as input strokes for the Geos block after marking the near-boundary-pixels as unclassified. The final segmentation mask is obtained from the Geos output.

IV. BILAYER/SUPERPIXEL AND GEOS/SUPERPIXEL SEGMENTATION

In this method, a late fusion is performed between the segmentations obtained by either Bilayer or Geos and the superpixel-based segmentation. In the following, we explain how this type of fusion is performed.

Assume P_i to be the i^{th} superpixel of the input image and is represented as the collection of the $p(x,y)$ which are assigned the same cluster C_i.

$$P_i = \{p(x,y) : p(x,y) \epsilon C_i\} \quad (1)$$

The final segmentation is obtained by assigning a unique label to every superpixel P_i: $L(P_i) = 1$ or 0 for foreground or background, respectively; according to the following equation:

$$L(P_i) = \begin{cases} 1, & if \ \sum_{\forall p(x,y) \in P_i} L(p(x,y)) \geq \frac{1}{2}|P_i| \\ 0, & Otherwise \end{cases} \quad (2)$$

where |.| is the cardinality operator. This fusion ensures assigning a unique label to superpixels. This helps a lot in getting rid of misclassified regions due to existence of inner edges in either the background or foreground.

V. BILAYER/GEOS/SUPERPIXEL SEGMENTATION

We perform the fusion between the three methods in two phases: the first phase is fusing both bilayer and Geos as described in section IV. The second phase is to take the output of the first phase and reclassify the pixels of the input image according to Eq. 2 for the superpixels of the input image.

VI. EVALUATION RESULTS

We have used our own re-implementation of the Godesic, Bilayer, and Superrpixel approach for the evaluation purposes. We applied the proposed approach on different benchmark video sequences []. We show visual results obtained for samples of our own video sequences only.

Figure 2 shows qualitative results of different types of the proposed fusion approaches. The results of individual Bilayer, Geodesic, or Superpixel segmentation show the performance of the traditional techniques for comparison purposes with the proposed approach, It is clear from the figure that fusion in general improves the segmentation quality, specially for the case of Bilayer/Geodesic/Superpixel fusion. For some few cases, the fusion results may be worse than those of the original methods. The main reason behind this is the superpixel clustering threshold, which is to be selected in a way that balances between the coarse and fine nature of the segmented objects. The threshold selection does not represent a big challenge given that it takes the same values for similar video streams. Also, the overall segmentation performance is not very sensitive to that threshold.

The charts in Figures 3-7 show quantitative results for the five video sequences of [2]. Error values are calculated against manually-labeled ground truth. As shown in the figures, fusion improves the accuracy of segmentation for most sequences with some efficiency cost. The worst cases occur when the error values of bilayer segmentation falls below the fusion values. Even for these cases, the fusion-obtained error values are very close to the minimum ones. The first two sequences are more difficult than the others, in terms of background variations []. Bilayer/Geodesic/Superpixel fusion gives the best results for these challenging cases. However, in sequences for which backgrounds are expected to be less challenging, Bilayer/Geodesic could be a good alternative.

VII. CONCLUSIONS AND FUTUR WORK

In this paper, we presented a novel video segmentation approach. The proposed approach depends on fusing a graph-based superpixel image segmentation algorithm with two of the state-of-the-art video foreground/background live separation in video streams, in four different ways. We proposed a merging algorithm to reduce the number of the generated superpixels for the operation of the fusion algorithm. The evaluation results showed that the proposed methodology improves the performance of the state-of-the-art in terms of accuracy. In the worst case conditions, for some "difficult" streams, the accuracy is almost preserved.

For future work, we will make investigations of improving the efficiency of the proposed approach to increase the supported frame rate.

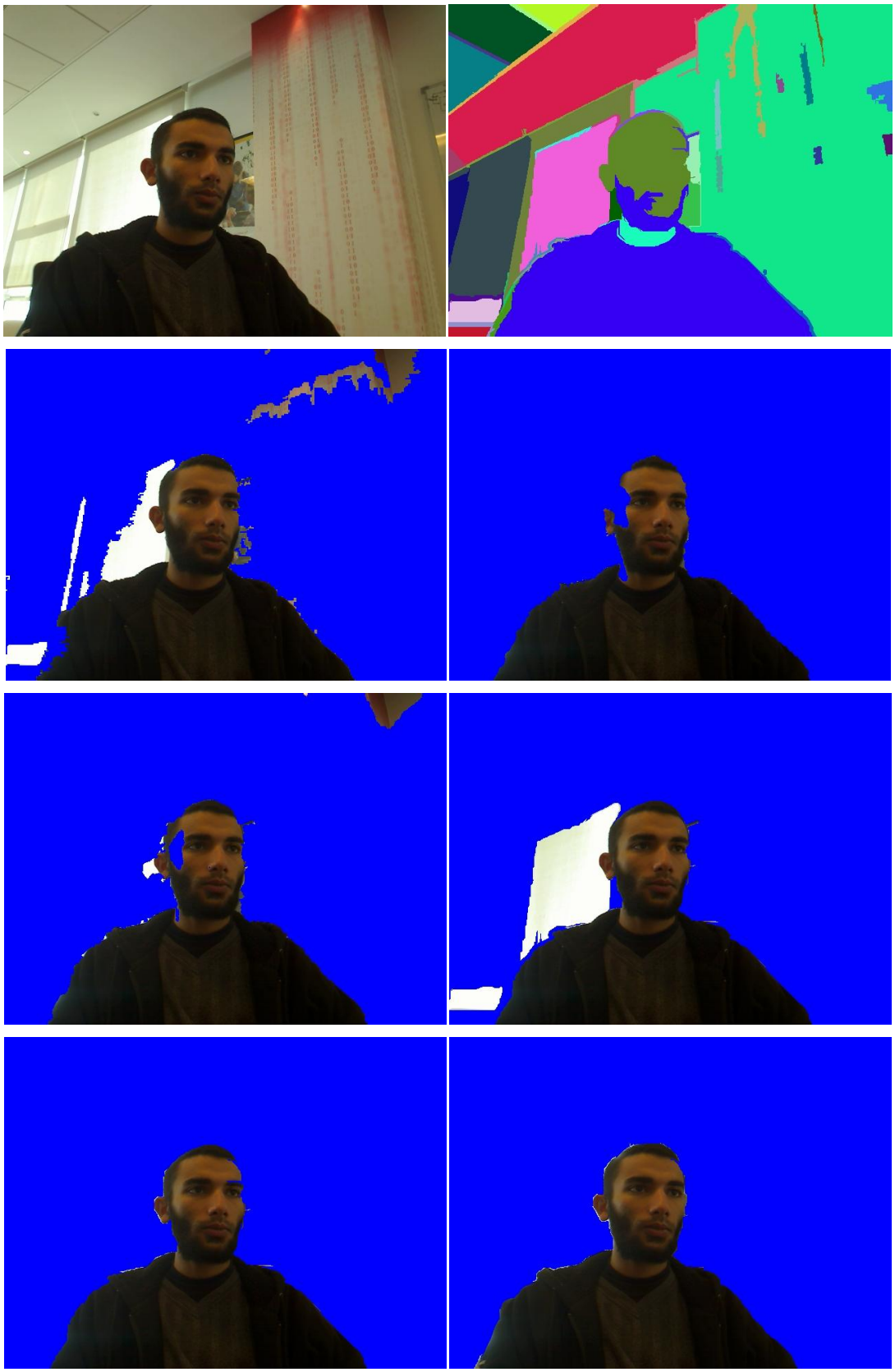
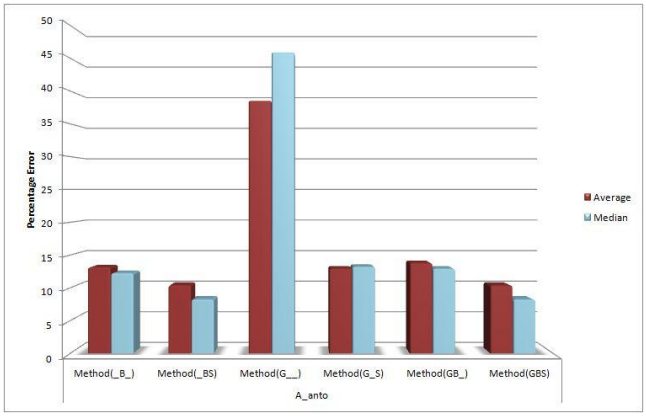
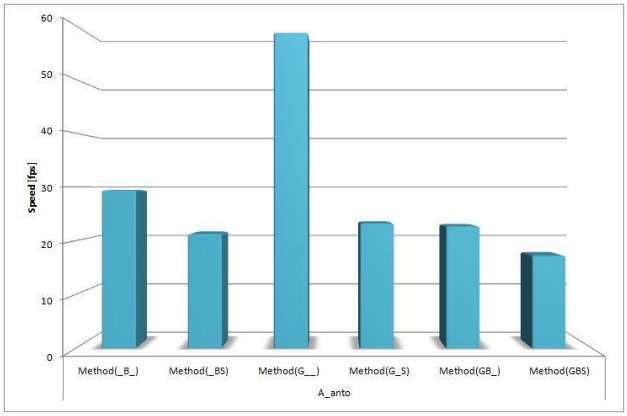


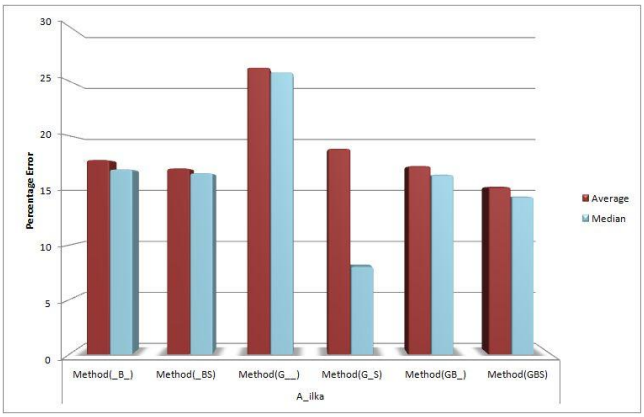
Figure 2: Qualitative segmentation results. From top left: The input frame, superpixel segmentation, bilayer segmentation, geodesic segmentation, fused bilayer/geodesic segmentation, fused bilayer/siperpixel fused segmentation, fused geodesic/superpixel segmentation, and segmentation results of fusion all of the three methods.

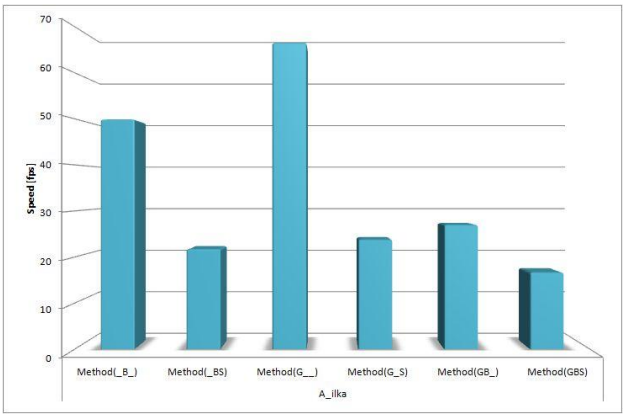




(a) Error values **(b) frame rate**

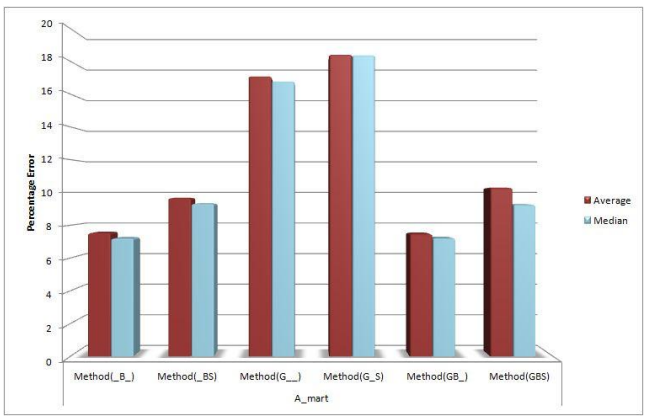
Figure 3: Evaluation results in terms of accuracy and speed for the Antonio sequence of [2]

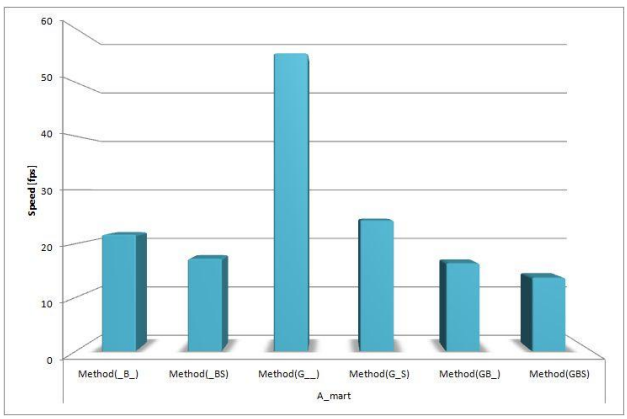




(a) Error values **(b) frame rate**

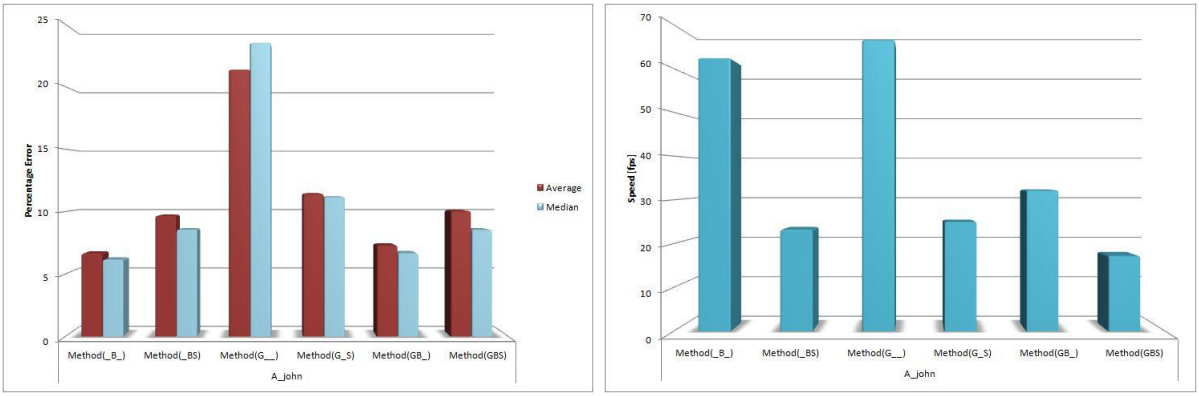
Figure 4: Evaluation results in terms of accuracy and speed for the Ilka sequence of [2]





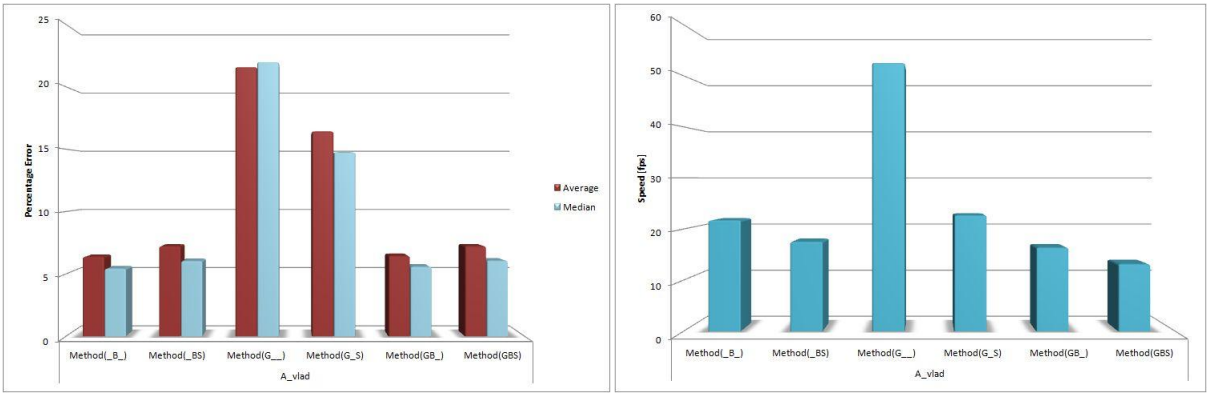
(a) Error values **(b) frame rate**

Figure 5: Evaluation results in terms of accuracy and speed for the Mart sequence of [2]



(a) Error values **(b) frame rate**

Figure 6: Evaluation results in terms of accuracy and speed for the John sequence of [2]



(a) Error values **(b) frame rate**

Figure 7: Evaluation results in terms of accuracy and speed for the Vlad sequence of [2]

I. REFERENCES

[1] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:359–374, 2001.

[2] A. Criminisi. *Database of monocular sequences labelled into foreground and background layers*. `http://research.microsoft.com/en-us/projects/i2i/data.aspx`.

[3] A. Criminisi, G. Cross, A. Blake, and V. Kolmogorov. Bilayer segmentation of live video. In *In: IEEE Conference on Computer Vision and Pattern Recognition*, pages 53–60, 2006.

[4] A. Criminisi, T. Sharp, and A. Blake. Geos: Geodesic image segmentation. In *ECCV '08 Proceedings of the 10Th European Conference on Computer Vision: Part I*, pages 99–112. Springer-Verlag, 2008.

[5] Z. Dong, L. Jiang, G. Zhang, Q. Wang, and H. Bao. Live video montage with a rotating camera. *Comput. Graph. Forum*, 28(7):1745–1753, 2009.

[6] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.

[7] V. Kolmogorov, A. Criminisi, A. Blake, G. Cross, and C. Rother. C.: Bilayer segmentation of binocular stereo video. In *In: IEEE CVPR (2005*, pages 407–414, 2005.

[8] K.-C. Lien and Y.-C. F. Wang. Automatic object extraction in single-concept videos. *Multimedia and Expo, IEEE International Conference on*, pages 1–6, 2011.

[9] A. Monnet, A. Mittal, N. Paragios, and V. Ramesh. Background modeling and subtraction of dynamic scenes. pages 1305–1312, 2003.

[10] J. Sun, W. Zhang, X. Tang, and H. yeung Shum. Background cut. In *In ECCV*, pages 628–641, 2006.

[11] J. Wang, P. Bhat, R. A. Colburn, M. Agrawala, and M. F. Cohen. Interactive video cutout. *ACM Trans. Graph.*, 24(3):585–594, July 2005.

[12] P. Yin, A. Criminisi, J. Winn, and I. Essa. Tree-based classifiers for bilayer video segmentation. In *In CVPR*, 2007.

[13] P. Yin, A. Criminisi, J. M. Winn, and I. A. Essa. Bilayer segmentation of webcam videos using tree-based classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(1):30–42, 2011.

[14] F. Zhong, X. Qin, and Q. Peng. Transductive segmentation of live video with non-stationary background. In *CVPR*, pages 2189– 196, 2010