RESEARCH ARTICLE OPEN ACCESS

Visual and Motion Saliency Using Unsupervised Block Based Approach for Video Object Extraction

A.Carolin Delviya1, Mrs.N.Mymoon Zuviria2
1PG Scholar, Department of CSE, National College of Engineering, Tamilnadu
2 Assistant Professor, Department of CSE, National College of Engineering, Tamilnadu
1caro.delvia@gmail.com
2rizrose1@yahoo.com

Abstract—
Video Object Extraction is the process of identifying and tracking the collection of image pixels that correspond to the projection of a real object in successive image planes of video sequence .This video object extraction fully based on their visual and motion saliency Induced features using Block-by-block wise model .A block-based method capable of dealing with noise, illumination variations, and dynamic backgrounds, while still obtaining smooth contours of foreground objects. In this Paper proposes Video Object Extraction (VOE) framework presents Saliency Information allows us to infer several visual and motion cues for learning foreground and background models in block by block basis. This framework aims to automatically segment the foreground object of interest from a video sequence without any user interaction or any training data. First the foreground and background objects are separated within video frames, which utilizes both visual and motion saliency information from an input video. A Conditional Random Field(CRF) is applied to effectively combine a saliency induced features which allows us to deal with unknown pose and scale variations of the foreground object (and its articulated parts).Experiments on a variety of videos verify that the proposed method is able to produce quantitatively and qualitatively satisfactory VOE results.
Keywords—— Video Object Extraction, Visual Saliency,Motion Saliency ,Conditional Random field, Segmentation

I. INTRODUCTION

The representation of video information, in terms of its content, is at the foundations of many multimedia applications such as broadcasting content-based information retrieval, video editing , activity recognition and its entertainment .Multimedia Application needs automatic techniques for extracting such objects from video data .This extraction techniques that enable the foreground objects from background objects. Our goal of image analysis to extract the meaningful entities from visual data.

A meaningful entity in image is object in real world such as a person, tree, and animals etc. video object is a collection of image pixels that corresponds to the projection of a real object in successive image planes of video object.

One of the fundamental and critical tasks in many computer-vision applications is the segmentation of foreground objects of interest from an image sequence.

Video object extraction (VOE) aims to segment foreground objects of interest from video data. The success of VOE helps machines understand the content of videos, and thus VOE is typically considered as a preprocessing technique for high level computer vision tasks such as human pose estimation ,event recognition, and video annotation.

Due to human vision , the eye can perhaps an video ,it can easily determine the subject or object of interest in an video. But in the Machine vision ,it cannot happen, because the object is presented in an unknown or messy background or the object cannot be clearly displayed in an video. These composite abilities captured by an human mind ,this process can be interpreted as Synchronized extraction of both foreground and background information from an video sequence.

Most of the researchers have been working toward closing the gap between the human vision and computer vision. However , without any prior knowledge on the object of interest or training data ,it is still very challenging for computer vision algorithms to automatically extract the foreground object of interest in a video. As a result, if one must to be design an algorithm to automatically extract the foreground from the video frame, quit a few responsibilities are must be deal.

Undefined object type with an amount of undefined object case in video frames.

Difficult or else unpredicted movement of objects appropriate to expressed element or random poses.

Uncertain form among foreground and background areas appropriate to parallel color, low disparity, deficient illumination, etc. conditions.
Conversely, if anyone can extort delegate information from foreground either background regions from a video. The extracted information can be exploited between regions, and therefore the assignment of object instance extraction can be addressed.
For videos captured by a static camera, extraction of foreground objects can be treated as a background subtraction problem. Most of the prior works either consider a fixed background or assume that the background exhibits dominant motion across video frames. These assumptions might not be practical for real world applications, since they cannot generalize well to videos captured by freely moving cameras with arbitrary movements.
Most of the prior work deals with foreground objects are extracted using pixel-by-pixel basis. Many Pixel-by-pixel Video Object Extraction framework do not capable dealing and dynamic Backgrounds. An inherent limitation of pixel– by–pixel processing is that rich contextual information is not taken into account. For example, pixel-based segmentation algorithms may require ad-hoc post-processing to deal with incorrectly classified and scattered pixels in the foreground mask.

In this paper we propose a Video Object Extraction (VOE) framework , which is able to extract the foreground object of interest from an video. This framework mainly focuses on the foreground object and based on the visual and Motion saliency induced features using Block-by-block model . Given such a video, we construct a compact color and shape model induced by motion cues, and extract the foreground and background color information accordingly. We integrate these feature models into a unified framework via a conditional random field (CRF),and this CRF can be applied to video object segmentation and further video editing and retrieval applications. One of the advantages of our method is that we do not require the prior knowledge of the object of interest, and thus no training data or predetermined object detectors are needed; All the feature models we utilize in our CRF are automatically extracted from the test input video in an unsupervised setting, and this cannot be easily achieved by most prior work.

Remainder of this paper This paper is organized as follows: Section 2 introduce related works on foreground object extraction and highlights the hand-outs of our method. Our proposed constructions are presented in sections 3 and 4. Finally section 5 concludes this paper.

II. RELATED WORK

In universal, one can address VOE troubles using supervised or unsupervised approaches. Supervised approaches need previous information on the object of attention and require that collect exercise data in advance for designing the related VOE algorithms.

Unsupervised approaches do not train any explicit object detectors or classifiers in advance. For the videos captured by a static camera, extraction of object instances can be treated as a background subtraction problem. In other words, objects can be perceived basically by subtracting the current frame from a video sequence [1], [2]. However, if the background is every time varying or is occluded by object instances, background modeling turn into exceptionally demanding assignment. For such cases, detectors usually mean at information the background model from the input video sequence frame, and the foreground object instance are considered as outliers to be perceive. For instance, an auto regression moving average model (ARMA) that approximate the natural exterior of self-motivated surface and area was proposed in [3], and it mostly compact with situation in which the background consists of normal view similar to sea waves or trees. Sun et al. [4] develop color gradients of the background to find out the limitations of the object instance. Some unsupervised advance aim at survey features related with the foreground object for VOE. For example, graph-based methods [5], [6] identify the foreground object regions by reduce the cost between neighboring unknown nodes/pixels information. Additional exclusively, one can segment the object instance by separating a graph addicted to displace pieces whose entire energy is reduced devoid of using any preparation information. Although moving consequences information in [5], [6], these approaches usually guess that the background/camera motion is central across video frames. For universal videos imprison by freely moving cameras, these technique may not simplify well. Different from graph-based methods, Leordeanu and Collins [7] proposed to monitor the co-occurrences of object features to identify the object instance in an unsupervised setting. Although hopeful consequences below pose, scale, occlusion, etc. variations were statement, their approach was only able to deal with rigid objects (like cars). Since Itti et al. [8] first derived the visual saliency of a single representation, several mechanism have been planned to extract the saliency information of images for the tasks of compression, classification, or segmentation. For exemplar ,Harding and Robertson [9] exhibit that the visual saliency can be develop to recover image density ratio by merge

SURF features and task-dependent previous information. Unlike density or categorization problems which might make the most of task or object category information for receive the related saliency, universal saliency detection or image segmentation tasks are resolve in an unsupervised setting. For example, standard spectrum analysis, Hou and Zhang [10] utilized the spectral residual as saliency information, while Guo et al.

The vast majority of the algorithms described in the literature belong to the pixel-by-pixel category. Notable examples include techniques based on modeling the distribution of pixel values at each location. For example, Stauffer and Grimson[11] modeled each pixel location by a Gaussian mixture model (GMM).

III. METHODOLOGIES

This section includes an overview of VOE framework. How foreground object is extracted from an video sequence. Fig .1 shows an illustration of process to extract the foreground object and its features. First the input video is given to the framework and video can be splited into number of frames called frame splitting.

Then we discuss the next part of the framework is described as the visual and Motion salient feature extraction. The system chosen one video frame ,we extract the visual salient features such as color, edge ,gradient etc. These features are given to the Visual saliency module ,in this we have obtain the visual saliency map. Before enter into the Motion saliency module, we estimate the motion sequence between the two video frames called optical flow.

Based on the optical flow map we have obtain the Motion saliency map. These two modules are main part of our framework because saliency information of both visual and Motion cues are extracted. After we extract the saliency features of both visual and Motion ,construct a compact color and shape model using Gaussian Mixture model(GMM).

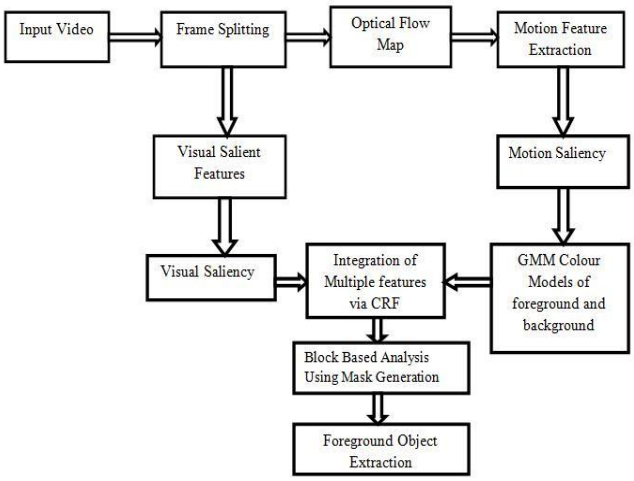

Fig.1 Foreground Object Extraction Framework

In this we classify the foreground and background color information and find the mean and covariance using Expectation –Maximization algorithm ..CRF is used to predict the label of an each observed pixel ,and we derive various energy function discussed in later section ,finally foreground object is extracted.

The first method named as visual saliency estimation , second Motion Induced shape cues, next learning color Information using Gaussian Mixture Model(GMM) and final Method describe the Integration Multiple Features using CRF.

A. Visual Saliency Estimation

Visual Saliency is the distinctive perceptual quality which makes some objects in real world ,which stand out from their neighbours. In this section includes two sub sections (i.e) first section deal with Region (superpixel) segmentation and the next describes Saliency estimation.

A.1. Region (super pixel) segmentation

To extract the visual saliency of each frame ,we perform Image segmentation on each video frame and extract color and contrast Information. Then region segmentation is performed on the first frame ,frame can be converted into the number of regions having the same shape ,color called Super pixel. The resulting frame segments (superpixel or regions) are applied to perform saliency detection. For example the following figure shows an region segmentations results. The first frame can be converted into 58 regions and find which region mostly having the same structure and color that region can be taken and perform region segmentation region can reduced into 16 region then the process will be repeated.

Then compute color contrast at the region level and the saliency for region(superpixel) is defined as the weighted sum of the region's contrast to all other regions in an image.The weights are set according to the spatial distances with farther regions being assigned smaller weights. Storing and calculating the regular matrix format histogram for each region is inefficient since each region typically contains the small number of colors in color histogram of the whole image.

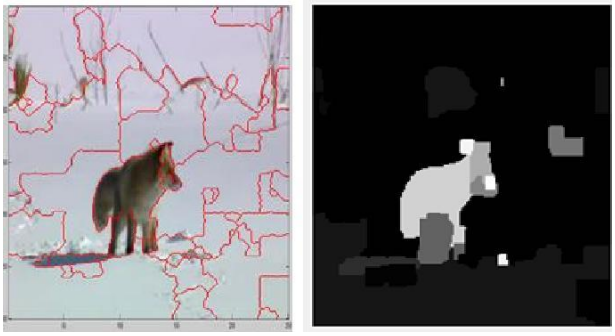

Fig.2 Region (superpixel) Segmentation on the first video frame 2a)frame can be converted into 16 region ,2b)Visual saliency Map.

$$S(r_k) = \sum_{r_k \neq r_i} \exp(D_s(r_k, r_i)/\sigma_s^2)\omega(r_i)D_r(r_k, r_i) \qquad (1)$$

For a region r_k,its saliency value is computed by measuring its color contrast to all other regions in the image ,where D_s is the Euclidean distance between the centroid of r_k and that surrounding super pixels or region r_i,while σ_s^2 controls the width of the kernel. The parameter $\omega(r_i)$ is the weight of the neighbor super pixel r_i,which is proportional to the number of pixel in r_i.The last term $D_r(r_k, r_i)$ measures the color difference between r_k and r_i.

The above figure 2a) the frame divided into the number of regions and compute saliency value of each region have the same color and shape etc.

Using this saliency estimation we have to compute color histogram for each region and obtain the visual saliency map. The figure 2b) shows an visual saliency map for fox video, and visual salient features such as color ,edge can extracted.

B. Extraction Of Motion-Induced Cues

Motion saliency helps to detect moving objects whose motion is discontinuous from its background. This section includes how to extract salient features of Moving objects in a video sequence. First motion saliency is derived from the optical flow map and extracts the color and shape information from the motion saliency results. For the motion saliency estimation first the moving pixel is detected and optical flow map for each frame can be derived.

B.1 Motion saliency Estimation

After obtaining the visual saliency map ,we extract the motion salient regions based on the retrieved optical flow map. Optical flow is the technique in which , estimate the motion between two frames at time t and t+1. To detect each moving part and its corresponding pixels ,perform the dense optical flow forward and backward propagation at each frame of a video. A Moving pixel at frame t is determined by

$$q_t = \hat{q_t}, t-1 \cap \hat{q_t}, t+1 \qquad (2)$$

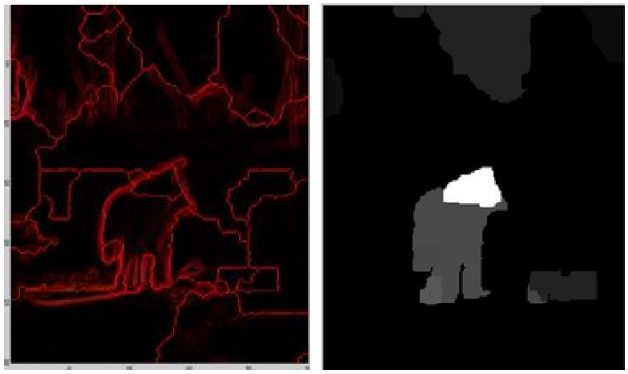

Fig (3a) Optical flow map (3b) Motion saliency estimation

Where ˆ denotes the pixel pair detected by forward or backward optical flow propagation. After determining the moving regions, we propose to derive the saliency score for each pixel interms of the associated optical flow information.

On the derived optical flow results to calculate the motion saliency M(i,t) for each pixel t at frame t, and the saliency score at each frame is normalized to the range [0,1].It is worth

Nothing that ,when the foreground object exhibits significant movements (compared to background) its motion will be easily captured, because the background regions are more contrast compared to the foreground ,at the moment the moving objects can be easily detected using the optical flow .

The above figure shows ,how to extract the motion salient regions in an video frame ,fig (3a) describes the optical flow because an video contains 24 frames in one minute, estimate the motion between 1st frame and second frame and compare the values of motion patterns and analyses all the frame ,then we easily obtain the optical flow map. Fig(3b) describes an Motion saliency map, it can be estimated from the optical flow results. In an Video

frame ,moving pixel is considered as fox head ,the part of an head can be white colored it can contain the range [255,255] ,and remaining region moving object shape can be estimated.

B.2 Learning of shape cues

Although motion saliency allows us to capture motion salient regions within and across video frames ,those regions might only correspond to the moving parts of foreground object within some time interval. Assume the foreground should near the high motion saliency region as the method ,the entire foreground object is not identified easily. since it is typically observed that each moving part of a foreground object forms a complete sampling of the entire object induced by motion cues for characterizing the
foreground object. In the above section only shape of the moving region of foreground object can be calculated..

B.3 Learning color cues

In this section we calculate the color information of Moving foreground Object. Gaussian Mixture Model can be used to learning the foreground and Background color from the shape Information.

Besides the motion-induced shape information, both foreground and background color information are extracted ,and it is used to improve the VOE performance. According to the observation and the assumption that each moving part of the foreground object forms a complete sampling of itself and the foreground or background color models are constructed based on visual or motion saliency detection results at each individual frame; otherwise, foreground object regions which are not salient in terms of visual or motion appearance will be considered as background, and the resulting color models will not be of sufficient discriminating capability.
From the equation (4) that is shape likelihood determine the candidate foreground FS sape and background BS sape regions. In other words, the color information of
pixels in FS sape for calculating theforeground color GMM are determined, and those BS sape in for deriving the background color GMM. Once these candidate foreground and background regions are determined and Gaussian mixture models (GMM) G^{cf} and G^{cb} are used to model the RGB distributions for each model. The parameters of GMM such as mean vectors and covariance matrices are determined using an expectation-maximization (EM) algorithm. Finally, both foreground and background color models are integrated with visual saliency and shape likelihood into a unified framework for VOE.

C. Integration of Multiple Features Via CRF

Utilizing an undirected graph, conditional random field (CRF) is a powerful technique to estimate the structural information (e.g. class label) of a set of variables with the associated observations. For video foreground object segmentation, CRF has been applied to predict the label of each observed pixel in an image .at each pixel i in a video frame is associated with observation Z_i, while the hidden node F_i indicates its corresponding label (i.e. foreground or background). In this framework, the label F_i is calculated by the observation Z_i, while the spatial coherence between this output and neighboring observations Z_j and labels F_j are simultaneously taken into consideration. Therefore, predicting the label of an observation node is equivalent to maximizing the following posterior probability function is

$$\rho(F|I,\varphi) \propto \exp\{-(\textstyle\sum_{i\epsilon I}(\varphi_i) + \sum_{i\epsilon I, j\epsilon Neighbor}(\varphi_{i,j}))\} \qquad (3)$$

The pairwise term describing the relationship between neighbouring pixels Z_i and Z_j, and that between their predicted output labels F_i and F_j. Note that the observation z can be represented by a particular feature, or a combination of multiple types of features. To solve a CRF optimization problem, one can convert the above problem into an energy minimization task, and the object energy function E of (3) can be derived as

$$E = -\log(\rho) \qquad (4)$$

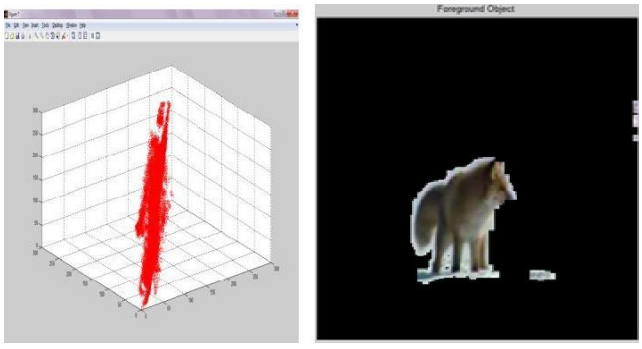

Fig (4a) Color models of foreground and background (4b)Extracted foreground object

In the same shot of a video, an object of interest can be considered as a compact space-time volume, which exhibits smooth changes in location, scale, and motion across frames. Therefore, how to preserve spatial and temporal consistency within the extracted foreground object regions across video frames is a major obstacle for VOE. Since there is no guarantee that combining multiple motion-

induced features would address the above problem, they need to enforce additional constraints in the CRF model in order to achieve this goal. In the proposed VOE framework multiple types of visual and motion salient features are utilized, and our experiments will confirm the effectiveness and robustness of our approach on a variety of real-world videos.

The above figure(4a) shows an color models of foreground and background ,it can be obtained by Expectation-Maximization algorithm using mean and covariance metrics of Gaussian Mixture Model(GMM).Then the foreground object(i.e) fox can be extracted in figure(4b).

D.Block based Analysis using Foreground Mask Generation

We exploit the overlapping nature of the block-based analysis to alleviate this inherent problem. Each pixel is classified as foreground only if a significant proportion of the blocks that contain that pixel are classified as foreground. In other words, a pixel that was misclassified a few times prior to mask generation can be classified correctly in the generated foreground mask. This decision strategy, effectively minimizes the number of errors in the output.

Formally, let the pixel located at (x, y) in image I be denoted as $I(x,y)$. Furthermore, let $B^{fg}_{(x,y)}$ be the number of blocks containing pixel (x, y) that were classified as foreground (fg), and $B^{total}_{(x,y)}$ be the total number of blocks containing pixel (x, y). We define the probability of foreground being present in $I(x,y)$ as

$$P(fg|I_{(x,y)}) = B^{fg}_{(x,y)} / B^{total}_{(x,y)} \qquad (5)$$

If $P(fg|I_{(x,y)}) \geq 0.90$ it is labeled as foreground. From the equation (5),each block can be identified by foreground and background .

IV. RESULT ANALYSIS

In this section, we conduct experiments on a variety of videos. We first verify the integration of multiple types of features for VOE, and show that it outperforms the use of a particular type of feature. We also compare our derived saliency maps and segmentation results to those produced by other saliency based or state-of-the-art supervised or unsupervised VOE methods. We consider the four videos named as fox, ant, bird, water-ski, each video derive the Motion Induced Shape Information and visual saliency Map, compare their Ground Truth Information.

To improve the VOE performance, we measure precision, recall metrics for Motion Induced shape Information and visual saliency based segmentation. For the following table shows the F-measure score for various videos and get the better average results. The F-measure defines measure of test accuracy. It considers both the precision and the recall of the test to compute the score .The F-measure can be interpreted as a weighted average of the precision and recall where an F-measure reaches its best.

$$F - measure = 2 \cdot \frac{Recall \cdot precision}{Recall + precision} \qquad (6)$$

Where precision, recall is given by fp, tp, fn, tn such as false negatives, true positives respectively. The higher F-measure value, then the foreground segmentation is more accurate. The following table shows the F-measure between visual saliency and motion Induced shape.

Table I
F-measure for various videos

Methods	Fox	Ant	Bird	Water ski
Visual Saliency	0.84109326	0.88165	0.916528	0.9277332
Motion Induced shape	0.810389	0.854937	0.898938	0.90478

F-Measure is calculated using visual saliency method, has the better performance than Motion Induced shape for different set of videos (Fox, Ant, Bird, Water ski).

V. CONCLUSION

The proposed method is an automatic VOE approach which utilizes multiple motion and visual saliency induced features, such as shape, foreground/background color models, to extract the foreground objects in videos. A CRF model is used to integrate the above features, and additional constraints are introduced into our CRF model for preserving both spatial continuity and temporal consistency, when performing VOE. Compared with state-of-the-art unsupervised VOE methods, the proposed approach is shown to better model the foreground object due to the fusion of multiple types of saliency-induced features. A major advantage of our proposed method is that neither the prior knowledge of the object of interest is required (i.e., the need to collect training data), nor the interaction from the users during the segmentation progress. Experiments on a variety of videos with highly articulated objects or complex background presented verified the effectiveness and robustness of our proposed method.

References

[1.] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu,"Global contrast based salient region detection," in Proc. IEEE Conf.Comput. Vis. Pattern Recognit., Jun. 2011, pp. 409

[2.] K. Fukuchi, K. Miyazato, A. Kimura, S. Takagi, and J. Yamato, "Saliency-based video segmentation with graph cuts and sequentially updated priors," in Proc. IEEE Int. Conf. Multimedia Expo, Jun.–Jul. 2009, pp. 638–641.

[3.] M. Gong and L. Cheng, "Foreground segmentation of live videos using locally competing 1SVMs," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2011, pp. 2105–2112.

[4.] M. Grundmann, V. Kwatra, M. Han, and I. Essa, "Efficient hierarchical graph based video segmentation," in Proc. IEEE Conf. Comput. Vis.Pattern Recognit., Jun. 2010, pp. 2141–2148.

[5.] Y. J. Lee, J. Kim, and K. Grauman , "Key-segments for video object segmentation," in Proc. IEEE Int. Conf. Comput. Vis., Nov. 2011,pp. 1995–2002.

[6.] M. Leordeanu and R. Collins, "Unsupervised learning of object features from video sequences," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2005, pp. 1142–1149.

[7.] A. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, and S. J.Dickinson,"TurboPixels: Fast superpixels using geometric flows," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 12, pp. 2290–2297, Dec. 2009.

[8.] K.-C. Lien and Y.-C. F. Wang, "Automatic object extraction in single concept videos," in Proc. IEEE Int. Conf. Multimedia Expo, Jul. 2011, pp. 1–6

[9.] Z. Lin and L. S. Davis, "Shape-based human detection and segmentation via hierarchical part-template matching," IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 4, pp. 604–618, Apr. 2010

[10.] F. Liu and M. Gleicher, "Learning color and locality cues for moving object detection and segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 320–327.

[11.] T. Matsuyama, T. Wada, H. Habe, and K. Tanahashi, "Background subtraction under varying illumination," Syst. Comput. Japan, vol. 37, no. 4, pp. 77–88, 2006.