A Novel Approach to Background Subtraction Using Visual Saliency Map

Soheil Tehranipour, Hamidreza Rashidy Kanan*

Department of Electrical, Biomedical and Mechatronic Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran

Received 13 March 2015; accepted 8 June 2015

Abstract

Generally human vision system searches for salient regions and movements in video scenes to lessen the search space and effort. Using visual saliency map for modelling gives important information for understanding in many applications. In this paper we present a simple method with low computation load using visual saliency map for background subtraction in video stream. The proposed technique is based on finding image segments whose intensity values can be distinguished accurate. The practical implementation uses a sliding window approach, where the distributions of the objects and surroundings are estimated using semi-local intensity histograms. This introduced method requires no training so it can be used in embedded systems like cameras due to low load in calculation. So with our background subtraction algorithm we can detect pre-defined targets. Also the automatically video regions detected by proposed model are consistent with the ground truth saliency maps of eye movement data. Comparisons with state-of-the-art background subtraction techniques indicate that the introduced approach results in high performance and accuracy.

Keywords: Background subtraction, Visual Saliency map, real-time platform, video processing.

1. Introduction

Background subtraction is a generally used real-time method for detection and extraction of foreground objects in a video file. It is the first remarkable step in most computer vision applications, including human-computer interaction, marketing and advertisement, traffic monitoring, and video surveillance. This leads improvement of background subtraction algorithms and a lot of post-processing methods with the goal of better and faster performance.

Building a model of the background is the most common technique for background subtraction and then Foreground objects as interest points are detected by calculating the difference between the background model and current frame of the scene. A binary foreground mask can be created by classifying a pixel with a difference more than predefined threshold as foreground object or interest point. Unfortunately, there are some factors which decrease accuracy of getting foreground masks. For example, a good algorithm for background subtraction should be robust to changing light and illumination in the scene, incorporating new objects into the background model

* Corresponding author. Email: h.rashidykanan@qiau.ac.ir

and able to ignore the movement of small background elements such as leaf of tree due to wind. Also all of these operations should be low cost computes which allows us to use it in real-time platforms.

On the other hand, another approach is using object detection techniques, which are trained to find specific object categories like persons, cars and etc. [1]. Reported results show [2, 3], the drawback which is need to wide training process although they are not able to categories objects that do not belong to the predefined categories. Also the training process itself is heavy and usually the final performance of the methods is given by the content of the training set.

In the continued way, a recent approach to background subtraction is carried by visual saliency map. In general saliency can be mapped with the attention mechanism of the human visual system in the real world, which enables human beings to quickly focus on salient and specific objects without training and then learning phase [4]. Visual saliency is studied in the computer vision field and the result comes out that most attractive elements for our eyes are color, high contrast [5]. When observing an image several seconds and several minutes, a human viewer can slowly scan multiple parts of interest, and different observers may see various things in the image. While video watchers during viewing a video have only a little time to see each frame.

Some of other approaches which have been proposed for visual saliency map span the contrast of the image areas to its surroundings, which is done for getting features such as histogram, gradient orientation and color. The methods [6, 7] with this technique, require high process and many variable parameters to train so these are a restriction for the practical use of them. Another type of visual saliency map is obtained with the spectrum of the image's Fourier transform [7]. This algorithm has only few parameters, easy to implement, and the computation cost is very low for small resolution images but the accuracy is not well enough as previous methods discussed.

In this paper we are going to present a new approach for visual saliency map detection with measuring a semi-local feature contrast that will be used in background subtraction. The technique applies a sliding window approach, where the window size defines approximately the scale of the goal objects as interest points. The saliency of a point in the window is estimated by analyzing the conditional probability of a pixel to be detected from the distribution estimated inside the window in contrast to the distribution of the surrounding area. The contribution is the saliency estimation on semi-local areas instead of pixel level. In this way histogram will be a good choice for estimations of the distributions that is done during the algorithms in visual saliency.

This paper is organized as follows. In Section II, first some of the works which is done related to the background subtraction are going to be discussed. In Section 3 main algorithm for visual saliency map and the usage in background subtraction is proposed. Experimental results and conclusion are presented as follows.

2. Related Works

As shown in Figure 1 In the field Background subtraction there are main steps, such as Background modeling, Background initialization and Foreground detection. Background modeling explains the kind of model which is used to represent the background. The easiest and most used method to model the background is to get a background image without any moving object. In some outdoor scenes, the background is not available, and can always be changed with different situation, like illumination and light changes.

A practical and efficient background subtraction algorithm must solve some major problems. First, it should be robust against changes in lighting changes and also illumination. Second, it must avoid detecting non-stationary background objects, such as moving leaves, grass, rain and shadows of the object in the scene. Background modeling techniques could be classified into two major and broad categories: Non-Predictive and Predictive Modeling. The Non-Predictive modeling tries to model the scene as a time series, and creates a dynamic model at each pixel, with the past viewed images, and utilizes the valiancy of perversion between the actual value and the predicted one, to classify the pixels as the foreground or background. The latter overlook the order of the input observations, and develop a statistical model, such as the probability density function at each pixel [6]. Many of the pixel based background subtraction techniques are presented after the MOG (Mixture-of-Gaussians).In those techniques the intensity of every pixel in the image is defined by a mixture of K Gaussian distributions to model the general behavior of image. After presenting that, it is time to know about the most popular one, the GMM (Gaussian Mixture Model) [1].This model analyze and observe pixels and their values in a time manner and then includes modeling the distribution of the values with a weighted mixture of Gaussians. This background pixel model is able to tackle with the multimodal nature of many practical situations and helps to get better final outputs while other background motions, such as tree leaves or branches, are faced [6].

A background subtraction technique must adapt to gradual illumination changes like changing time and light of the day, motion changes, camera oscillations, high frequency objects like tree leaves. A few applications need to be embedded in the camera, so that the computational load and working in a real-time platform becomes the main concern. For the surveillance of outside scenes, robustness not in favor of noise and adaptive to lighting changes are also necessary. Nearly all the techniques act on each pixel independently. These techniques transfer entirely to post-processing algorithms the assignment of adding a few forms of spatial consistency to their results. By contrast, the method described is based on the guess that neighboring blocks of background pixels must follow similar variations over time. Whereas this guess holds most of the time, especially for pixels belonging to the same background object, it becomes difficult for neighboring pixels located at the border of several background objects. A block of a new video frame is classified as the background, if its observed image pattern is close to its reconstructions, using the PCA projection coefficients of 8-neighboring blocks. Such a technique is also described, but it lacks an update mechanism to adapt the block models over time. [1]

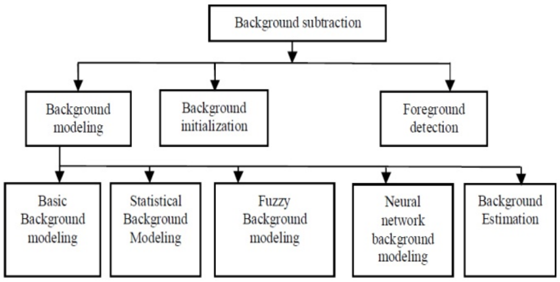
Background subtraction
Background modeling
Background initialization
Foreground detection
Basic Background modeling
Statistical Background Modeling
Fuzzy Background modeling
Neural network background modeling
Background Estimation

Fig. 1. Classification of background subtraction methods[7]

After presenting ICA (Independent Component Analysis) as an efficient algorithm that attempts to decompose a signal into independent non-Gaussian module, a series of images from a training phase, is described in the training of an ICA model [3] and then A two-level method based on a classifier is designed. A classifier first determines whether an image block belongs to the background or not. Classification algorithms are also the reason for getting through other algorithms, where the background model learns its patterns by self-training with artificial fuzzy and neural networks. Presently available algorithms for extracting the background from frames deal only with non-moving objects in a frame [7]. So far there is no algorithm that background model administrates moving objects in a frame. In many algorithms for background subtraction, the reference image is only a guess and in most of them the shadow effect is not discussed. On the other hand working a real-time platform can be a contribution and also a major filter to shine in all these methods above.

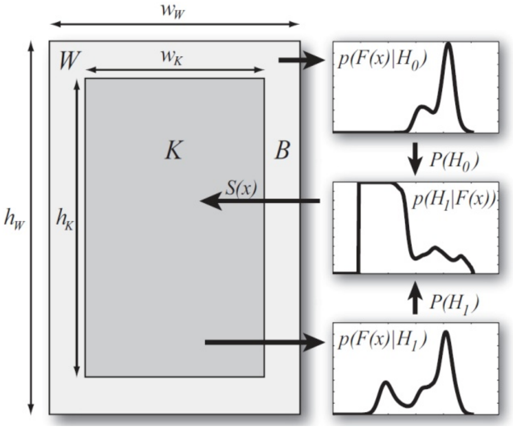

Fig. 2. Illustration of the definition of the saliency measure

3. Proposed Method

As mentioned above, in this paper finding background with visual saliency map is our goal. So we begin by introducing the general definition of our visual saliency map measure, and then we will go through the methodology and how to apply is in a video stream.

3.1. Visual saliency map measure

onsider a rectangular window like W in the image that is our purpose to detect salient point and edges, which is separated into two parts:

- one section is inner part : known as K,
- the other section is border : known as B,

W,K and B are shown in Figure 2 Here we define Width and height of W and K as w_W, h_W, w_k, and h_K and x $\in R^2$, that are also our parameters in source code with these names. These points are inside W, and F(x) is some feature value which will be calculated. Also we use image HSV instead of RGB of the image channel values as features F, because of the application of HSV model. With HSV we can detect illumination and intensity so in future, it is more robust to use it for pre-defined purposes. In this algorithm we have two assumptions,

- H_0: for points which are not salient,
- H_1: for points which are salient,

Here as initial hypothesis, H_1 is valid for the points which are located inside K, and H_0 is valid for points that are located inside the B. with this note, we can estimate the conditional feature distributions P(F(x)| H_1) and P(F(x)| H_0) from the feature values in K and B. Then according to Bayes' theorem P (A|B) = P (B|A) * P (A)/P (B) we can define

$$\mathrm{P}(\mathrm{F(x)}|\, H_1) \;= \frac{\mathrm{P}(\mathrm{F(x)}|\, H_1)\mathrm{P}(H_1)}{\mathrm{P}(\mathrm{F(x)})} \qquad (1)$$

$$\mathrm{P}\big(\mathrm{F(x)}\big) = \mathrm{p}(\mathrm{F(x)}|\, H_1)\mathrm{p}(H_1) + \mathrm{P}(\mathrm{F(x)}|\, H_0)\mathrm{P}(H_0) \qquad (2)$$

$$\mathrm{P}(H_1|\mathrm{F(x)}) \;= \frac{\mathrm{P}(\mathrm{F(x)}|\, H_1)\mathrm{P}(H_1)}{\mathrm{p}(\mathrm{F(x)}|\, H_1)\mathrm{p}(H_1) + \mathrm{P}(\mathrm{F(x)}|\, H_0)\mathrm{P}(H_0)} \qquad (3)$$

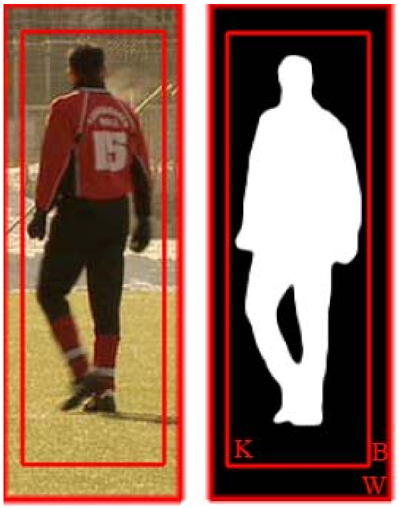

Fig. 3. Example of saliency values for a single window (W and K marked with red border). The features F used are the image intensity values.

In following steps of the algorithm now we can estimate the probability of H_1 for each of the points in K with calculated $\mathrm{P}\big(H_1\big|\mathrm{F(x)}\big)$. Saliency measure as S(x) for a point x in K to be the estimated probability:

$$\mathrm{S(x)} = \mathrm{P}(H_1|\mathrm{F(x)}) \qquad (4)$$

Figure 3 shows a sample of the evaluated values of S(x) with intensity as features F. The defined visual saliency measure S(x) gives out the contrast of the feature values between the B and K. Although visual saliency measure is related to other feature measures like the features that is discussed by Boiman and colleagues [8], but the main problem is that a posterior probability model is defined and will do the evaluation in semi-local windows, instead of one pixel at a time. By considering larger and bigger windows the algorithm will be able to be more independent to any model for the probability distributions.

3.2. Visual saliency map detector

In this section we will show that how it is applicable to use the visual saliency measure. Here I(x) is an image (this image will be used for background extraction using visual saliency map) with height h_I and width w_I. Also W(i) will be the window at center of the pixel with location i. The window W(i) will be moved as sliding all over the image I with a step s_W, and then calculation the S_i(x) as the saliency measure would be very easy just with a formula. In the final step of this section, to get a map of a pixel, the sampling step s_W is selected so the windows do overlap.

$$S(x) = \max_j\{S_j(\mathrm{x})|\mathrm{x} \in W_j\} \qquad (5)$$

Figure 4 shows the computation of the pixel saliency values.

According to Boiman [8] also prior probabilities P(H_1) and P(H_0) = 1 -P(H_1) would be required. A canonical choice would be to use the number of pixels in K and B.

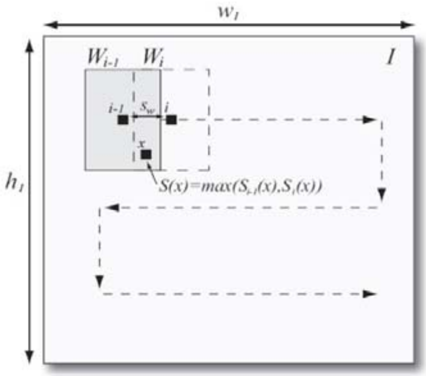

Fig. 4. Design of the computation of the saliency value for image I

Although based on our trial and error during the algorithm, we know that is better to use a lower value for $P(H_1)$ so here due to get more suitable output for next steps, $P(H_1)$ will be set to 0.20 in all next steps. Some threshold should be considered In order to fragment the salient target to more details from the non-salient background; we call it threshold of visual saliency S(x). The threshold “t” corresponds to the lowest probability we use a salient pixel to have for a hypotheses H_1, and we set it empirically to be t = 0.6 for still images and 0.66 for video sequences. Also another suggestion has been proposed that using morphological operations would be effective to better result of salient targets. Next all 4-connected sets which cover less than 0.15 percent of the image area will be removed and perform morphological closing with radius 0.15.min $\{h_I , w_I\}$.

Now we want to calculate the possibility so $\mathrm{h_K(F)}$ and $\mathrm{h_B(F)}$ are considered as the histograms of K as the core of the image for background subtraction and B as the border of considered part of the image as shown in Figure 4 Also g(F) will be function for the Gaussian smoothing used for detemination of the edge and interest points boundaries. Below the formula for estimating probability density of $\hat{\mathrm{p}}(\mathrm{F(x)}|\ \mathrm{H_1})$ and $\hat{\mathrm{p}}(\mathrm{F(x)}|\ \mathrm{H_0})$ are given:

$$\hat{p}(\mathrm{F(x)}|\ H_0) = \mathrm{N}(g(F) * h_B(\mathrm{F})) \qquad (6)$$

$$\hat{p}(\mathrm{F(x)}|\ H_1) = \mathrm{N}(g(F) * h_K(\mathrm{F})) \qquad (7)$$

Where N is a normalization operation:

$$\mathrm{N}\big(\mathrm{f(x)}\big) = \frac{1}{\sum_x f(x)} f(x) \qquad (8)$$

In the function above, $\hat{p}(\mathrm{F(x)}|\, H_i)$ is the estimate computed using only the frame K as the kernel. $\hat{p}(\mathrm{F(x)}|\, H_i)$ is the density estimate for a frame K that has been filtered out, and $0<\gamma<1$ will be a pre-defined parameter for making the exponentially degenerating effects of last frames in the video scene. The recursive estimator keeps only the previous value in memory and will bring only a marginal addition to the computational load.

In the next step we want to calculate the histograms h_K(F) and h_B(F). In this way we need to traverse over the pixels only one time to get the histograms for every window. When applying the proposed algorithm for video sequence, a simple kind of the process is to estimate each frame individually. However most often in real video scenes the changes within a small number of frames are very small.[12] We can use this fact by applying a fading memory beside a recursive least square estimator to filter $\hat{p}(\mathrm{F(x)}|\, H_1)$ and $\hat{p}(\mathrm{F(x)}|\, H_0)$. The estimator feeds as fast as it can to form the following equation.

$$\overline{p_k}(\mathrm{F(x)}|\, H_i) = \gamma\, \overline{p_{k-1}}(\mathrm{F(x)}|\, H_i) + (1-\gamma)\, \overline{p_k}(\mathrm{F(x)}|\, H_i) \qquad (9)$$

After getting through visual saliency map, the post processing step is very vital in order to remove small pixels which have been separate and to omit some small holes in the objects. Here morphological algorithms such as closing and erosion will be very effective due to their low computational load and ability to work in real-time platform. Fig. 5. is an example of the visual saliency map of a video stream from London’s street database. The performance of the proposed visual saliency map for with the approach of background subtraction is ordered by the size of the window W which has been described in previous section. Since the interest points in every frame of the video may appear in several scales it is suggested that a measurement tool would be used with different window sizes. The saliency value of the pixel will be maximum in overall scales. In order to calculate S(x) we need to estimate the conditional probability density functions p(F(x)| H_1) and p(F(x)| H_0). We do this by computing normalized feature histograms in the K as the core of image and in B as the border. For robustness we will smooth the histograms using Gaussian. The features F(x) used in our trial are the image intensities from HSV values. In the case of HSV values as features F(x), we compute the measure independently for each channel and take the maximum as the final saliency value. Although the scale is an important parameter, the measurement tool is not performing well against changes in the window size.

4. Experimental Results and Discussion

In the experimental section of this article, we will get the proposed visual saliency map for background subtraction method using the public video datasets Water Surface and London’s street. Experiments are performed on a collection of test images that are originally used in [6] and [14]. The saliency maps for the comparison mathods are computed using the programs give and for our method using our own implementation with default parameters.

True positive (TP) for a correctly classified foreground pixel, true negative (TN) for a correctly classified background pixel, false positive (FP) for a background pixel that was incorrectly classified as foreground and false negative (FN) for a foreground pixel that was incorrectly classified as background were calculated for each pixel in background subtraction method.

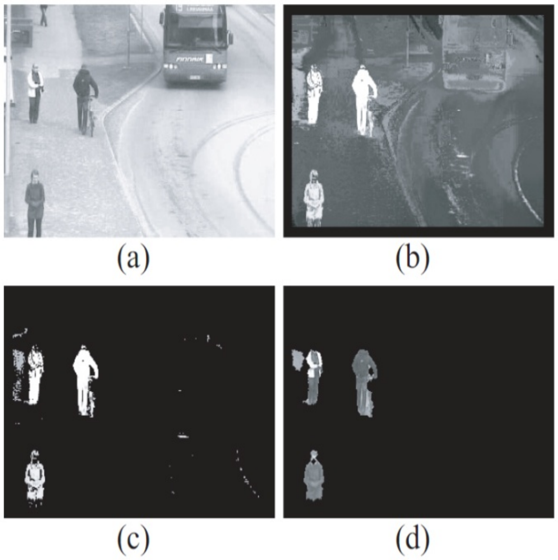

Fig. 5. Example of saliency detection, using single scale and image intensity values as features. (a) Original image. (b) Saliency values. (c) Threshold saliency values. (d) Final segmentation result.

After the classification of every pixel into one of those four groups had been done, sensitivity, precision, F1 and similarity were calculated with equations below. Sensitivity (Recall) measures the section of actual positives which are correctly identified. Precision is used to describe and measure the estimate or predict. Recall, also known as detection rate, gives the percentage of detected true positives as compared to the total number of true positives in the ground truth where is the total number of true positives in the ground truth. Moreover, we considered F1 that is the weighted harmonic mean of precision and recall. [13]

$$\textit{Sensitivity or Recall or TP Rate} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{10}$$

$$\textit{Precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{11}$$

$$\textit{Similarity} = \frac{\mathrm{TP}}{\mathrm{TP+FP+FN}} \tag{12}$$

$$\textit{F1} = \frac{2(\mathrm{Recall})(\mathrm{Precision})}{\mathrm{Recall+Precision}} \tag{13}$$

The proposed visual saliency map for background subtraction is evaluated in three phase with variables below. According to last section, we can make it possible to analyze the saliency in both grayscale and color images so the feature “F” are the intensity values in the case of grayscale image and the HSV channel values in the case of color images.

$$w_W = [0.2,0.4,0.6].\max\{w_I, h_I\}$$
$$h_W = [0.35,0.4,0.6].\max\{w_I, h_I\}$$
$$w_K = [0.1,0.3,0.5].\max\{w_I, h_I\}$$
$$h_W = [0.25,0.3,0.5].\max\{w_I, h_I\}$$
$$s_W = 0.01.\max\{w_I, h_I\}$$

At last in order to have a fair comparison with the method presented in [1], we also apply the removal of small holes and morphological closing to their results.

Finally we did another set of experiments using another video stream as Water Surface. The results are compared with the methods in [10, 12, 13 and 14], and the results are shown in Table 1. From the saliency measurement tool Fourier transform approach [6] has a specific version for main video sequences and the visual saliency which is introduced in [10] is applied to each frame separately. With the background extraction method of [4] we use grayscale frames and default parameters, but in order to get results faster and working in in real-time platform with dynamic video scenes, the learning phase is skipped and has been left to future works. Also it is applied to each frame separately with fading parameter, that has been set to = 0.8. The final test has an additional video illustrating non stationary camera.

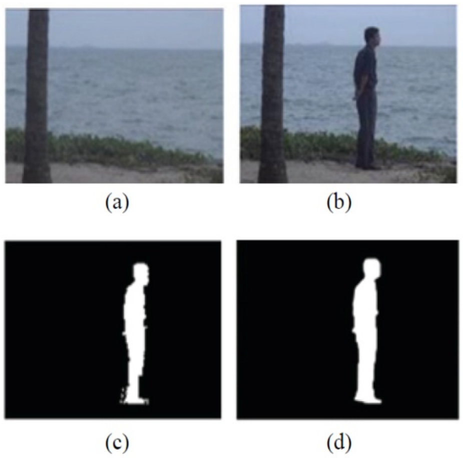

Fig. 6. Results with Water Surface test Sequences (a) Background reference image (b) Current frame (c) Proposed method (d) Ground truth method.

The scenes in the videos of [13] are highly dynamic and sharp, and very difficult for previous background subtraction methods. As you can see, our method still cannot compete with Mau's method [9] and Cheng's method [14] due to its lack of learning step but better than Tong's method [12] and Klare's method [13].

Table 1

Final Results with Water Surface test Sequences

method	Recall	Precision	Similarity	F1
Tong's method [12]	0.566	0.7078	0.7682	0.6290
Klare's method [13]	0.8835	0.6497	0.5872	0.7488
Guraya's method [10]	0.7615	0.8463	0.7434	0.8017
Cheng's method [14]	0.8026	0.9637	0.9048	0.8758
Mau's method [9]	0.9216	0.8512	0.9183	0.8850
Our Novel Method	0.8227	0.8369	0.8093	0.8297

5. Conclusion

This paper proposes a novel background subtraction method using visual saliency map model for video stream. The proposed technique is based on finding image segments whose intensity values can be distinguished accurate. The practical implementation uses a sliding window approach, where the distributions of the objects and surroundings are estimated using semi-local intensity histograms. Experimental results show that we can use this approach for practical applications. Also a great tool for usage in real-time platform with videos. Future work will be explored from two aspects. First, we will investigate how to explore our model to other real world applications. The second direction is to propose novel background subtraction model that use fuzzy neural system to add learning ability to the system presented.

References

[1] O. Barnich, M. V. Droogenbroeck. "ViBe: A universal background subtraction algorithm for video sequences." Image Processing, IEEE Transactions on 20.6 (2011): 1709-1724.

[2] S. Choudri, J. M. Ferryman, and A. Badii. "Robust background model for pixel based people counting using a single uncalibrated camera." Performance Evaluation of Tracking and Surveillance (PETS-Winter), 2009 Twelfth IEEE International Workshop on. IEEE, 2009.

[3] L. Maddalena, and A. Petrosino. "A self-organizing approach to background subtraction for visual surveillance applications." Image Processing, IEEE Transactions on 17.7 (2008): 1168-1177

[4] S. Lemercier, A. Jelic, R. Kulpa, "Realistic following behaviors for crowd simulation," EUROGRAPHICS 2012, Vol. 31, No. 2, 2012.

[5] K. Kim, T.H. Chalidabhongse, D. Harwood, L.S. Davis. "Real-time foreground background segmentation using codebook model," Real-Time Imaging, 11(3):172–185, June 2005.

[6] J. T. Jose, V.K. Govindan, "Efficient algorithm for varying area based shadow detection in video sequences," International Journal of Computer Applications (0975 – 8887), Volume 72– No.16, June 2013.

[7] T. Bouwmans, F. El Baf and V. B. Statistical Background, Modeling for Foreground Detection: A Survey, volume 4 of Handbook of Pattern Recognition and Computer Vision, chapter 3. World Scientific Publishing, 2010.

[8] O. Boiman and M. Irani, "Detecting Irregularities using saliency map in Images and in Video", ICCV, 2015.

[9] X. Xie Mau, W-Y. Ma, H.J. Zhang and H-Q. Zhou, Image Adaptation Based on Attention Model for Small-Form-Factor Devices", ICMM, 2013.

[10] F. F. Elahi Guraya, et al. "A non-reference perceptual quality metric based on visual attention model for videos." Information Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference on. IEEE, 2010.

[11] A. Shariq Imran, F. F. Elahi Guraya, and F. Alaya Cheikh. "A visual attention based reference free perceptual quality metric." Visual Information Processing (EUVIP), 2010 2nd European Workshop on. IEEE, 2010.

[12] Y. Tong, F. Alaya Cheikh, A. Tremeau and H. Konick, "Full Reference Image Quality Assessment Based on Saliency Map Analysis," Accepted for publication in the International Journal of Imaging Systems and Technology, in 2013.

[13] Y. Klare, F. Alaya Shci, F. F. E. Guraya and A. Tremeau, "A Visual Saliency Model for Perception-based," Accepted toVisual Communications and Image Processing VCIP 2014, China.

[14] Z. Cheng and Q. Li, "Video quality assessment using a saliency model of human visual speed perception," Journal of the Optical Society of America A 24, B61-B69 (2009).