Identification of Images in Motion Fields using Saliency

Ms. T. Chindrella Priyadharshini
Assistant Professor,
R.M.K College of Engineering and Technology, Puduvoyal, India.

Ms. P. DurgaDevi,
Research Scholar,
Anna University, Chennai, India.

Ms. M. Swarnalatha
Assistant Professor,
R.M.K College of Engineering and Technology, Puduvoyal, India.

***Abstract*— Automatically detecting people in videos is the first step in a wide range of tracking applications, which can be successively used in video surveillance systems, traffic monitoring, Object Detection and Action Recognition systems. Therefore detecting the outstanding feature of an image is considered to be more important. The outstanding feature of an image with respect to its neighborhood pixel is termed as "Saliency". In this proposed work, a Locally Adaptive Regression Kernel based object detection is employed for detection process. The proposed method is found to be simple and effective since the process does not require prior knowledge about objects in videos and also the model is parameter independent fast process.**

Keywords—Saliency detection, Motion fields, Regression kernel, Similarity Matrix

I. INTRODUCTION

Human Visual System (HVS) is perceptually more sensitive to certain colors, intensities. Computer vision is a field that includes methodologies for acquiring, processing, examining and empathizing images from the environment. In general, high volume of data obtained from the surrounding is processed in order to generate appropriate results in the form of decisions. It is also related with the theory behind artificial systems that extract information from images. The image may be in any formats such as video sequences, views from many cameras. The subdivisions of computer vision includes event detection, action recognition, motion estimation and image restoration. Saliency detection is the major operation involved in these applications. Saliency typically arises when there is a contrast exists between items and their neighborhood. The information captured by the human eye is very vast than that the central nervous system can process. Saliency detection is an emerging and interesting process in video processing applications. In the past decade, a number of Visual Saliency Models (VSM) have been proposed to simulate the behavior of eyes such as Saliency Tool Box (STB), Neuromorphic Vision Toolkit (NVT) and widely used for salient object detection and segmentation, tracking, image and video compression applications. but they require high computational cost and their remarkable results mostly rely on the choice of parameters. Various obstructions such as bright background, sudden illumination makes the process more complex. Most conventional object detectors require training in order to detect certain object categories[1]. But human vision can focus on general salient objects rapidly in a clustered visual scene without training because of the existence of visual attention mechanism. So human can easily address with universal object detection well, which is becoming an fascinating subject for progressive researches.

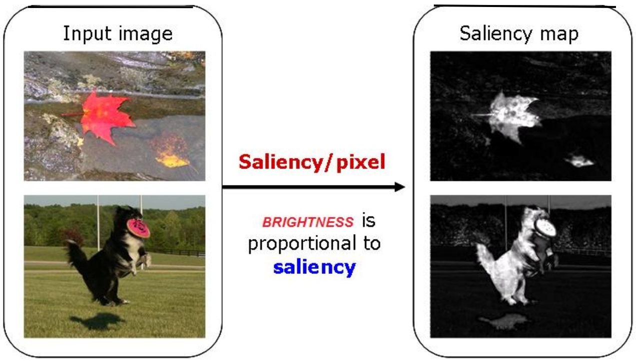

Fig 1. Representation of Saliency

Different from conventional segmentation problem of separating the whole scene into discrete parts, saliency detection directs at finding semantic regions and filtering out the unimportant area.

A. LIMITATIONS OF THE EXISTING MODELS

Tresiman [3] proposed the famous feature integration theory (FIT) which described visual attention as having two stages. A set of basic visual countenances, such as color, gestures and edges, is processed in parallel at the preattentive stage. And then, in the restricted-capacity process stage, the visual cortex executes other more complex operations like face recognition and others [4]. A master map or a saliency map [5] is computed to indicate the locations of salient areas. Distinctive features (e.g., luminous color, high velocity motion, and others) will "pop out" automatically in the preattentive stage, and then the salient areas become the object candidates. Various computational frameworks have been proposed to simulate human's visual attention, which are based on the bottom-up computational framework. Itti et al. proposed a bottom-up model and built a system called Neuromorphic Vision C++ Toolkit (NVT) [6]. Afterwards, following Rensink's theory [7], Walther extended this model to attend to proto object regions and created Saliency Tool Box (STB) [8]. He also implemented the model to accomplish object recognition tasks [9]. However, the high computational cost and the choice of parameters are still the weaknesses of these models. The spectral residual (SR) approach based on Fourier Transform was proposed by [10], which does not rely on the parameters and can detect salient objects rapidly. Later, Guo et al. [11] manifested the fact that the phase spectrum is key to calculating the saliency map and proposed a model called phase spectrum of Fourier transform (PFT) for saliency detection. Besides these models, Bruce et al. proposed a model of bottom-up overt attention based on the principle of maximizing information sampled from a scene [12]. Gao et al. presented the discriminant saliency detection model which requires a discriminant saliency selection process at first (training stage), and then the saliency map can be computed by the selected features at the testing stage [13]. A graph-based visual saliency detection was proposed in 2006 [14], which can powerfully predict human fixations but demands very high computational cost; [15] and [16] proposed the region-based approaches to calculate the feature maps for their saliency models, which perform a clustering at first and compute the feature maps by these clusters to reduce computational complexity. However, their models need to set many parameters to obtain useful results, and still can't work in real time (only a few frames per second). All these models mentioned above, however, only consider static images. Some function has been employed to add the motion feature to these models [17] in order to perform some applications. However, the additional motion channel will increase the computational cost of the model.

II. RELATED WORK

Incorporating motion into a saliency model without dramatically influencing its computational cost is a challenging task that motivates to develop a novel model to generate saliency map with the help of locally adaptive regression kernel. Moreover, besides SR and the proposed model, other models require tremendous computational cost and cannot meet real-time requirements on a personal computer (PC). Therefore, how to develop a saliency model that can work in real time is another consideration of this work. The problem of interest addressed in this proposed method is bottom-up saliency which can be described as follows: Given an image, we are interested in accurately detecting salient objects from the image without any background knowledge. In order to do this, in this proposed method local steering kernels is treated as features which capture local data structure exceedingly well. This approach is motivated by a Bayesian probabilistic framework, which is based on an independent parameter estimate of the likelihood of saliency and in turn leads to the local calculation of a "saliency map", which measures the similarity of a feature matrix at a pixel of interest to its neighboring feature matrices.

A. OVERVIEW OF THE PROPOSED APPROACH

In this proposed method saliency detection task is carried out as two-fold . First the local regression kernels as features is used which capture the underlying local structure of the data exceedingly well, even in the presence of significant distortions. Second a parameter independent kernel density estimation for such features, is used which results in a saliency map consisting of local measure, indicating likelihood of saliency. The archetype motivation behind these augmentations is the earlier work on adaptive kernel regression for image reconstruction and nonparametric object detection.

B. CONTRIBUTION OF LOCAL STEERING KERNEL

The key idea behind local steering kernel is to robustly obtain the local structure of images by analyzing the radiometric (pixel value) differences based on estimated gradients, and use this structure information to determine the shape and size of a canonical kernel. The local steering kernel is modeled as,

$$K(X_l - X_i) = \frac{\sqrt{\det(C_l)}}{h^2}\exp\left\{\frac{(X_l - X_i)^T C_l (X_l - X_i)}{-2h^2}\right\} \qquad (1)$$

where l ∈ {1,….. , P}, P is the number of pixels in a local window, h is a global smoothing parameter, and the matrix C_l is a covariance matrix estimated from a collection of spatial gradient vectors within the local analysis window around a sampling position $X_l = [x_1, x_2]_l^T$. The local steering kernel function $K(X_l - X_i)$ is calculated at every pixel location and normalized as follows,

$$W(X_l - X_i) = \frac{K(X_l - X_i)}{\sum_{l=1}^{P} K(X_l - X_i)} \quad , \text{i=1,…….,M} \qquad (2)$$

LSK reliably captures local data structures even in complex texture regions or in the presence of moderate levels of noise. Normalization of this kernel function yields invariance to brightness change and robustness to contrast change. From a human perception standpoint [18] the local image features are salient when they are distinguishable from the background. Computationally, measuring

saliency requires, the estimation of local feature dispersions in an image. For this purpose, a generalized Gaussian distribution is often employed . However, LSK features follow a power-law distribution (a long-tail dispersion). In other words, the LSK features are scattered out in a high dimensional feature space, and thus there basically exists no dense cluster in the feature space. Instead of using a generalized Gaussian distribution, a locally adaptive kernel density estimation method is used in the proposed method.

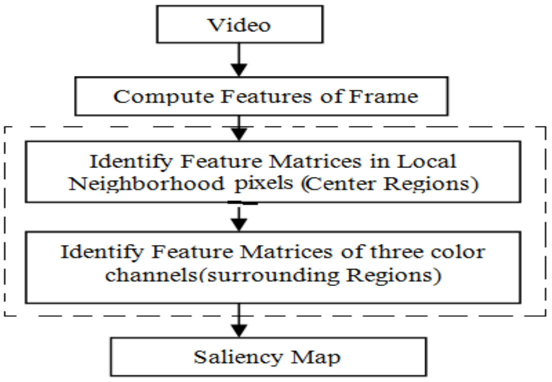
Video
Compute Features of Frame
Identify Feature Matrices in Local Neighborhood pixels (Center Regions)
Identify Feature Matrices of three color channels(surrounding Regions)
Saliency Map

Fig 2. Overview of the proposed Method

C. Managing color images

Up to now, saliency detection in a grayscale image is only used. If a color input image is employed, an approach to integrate saliency information from all color channels is needed. To avoid some drawbacks of earlier methods, instead of combining saliency maps from each color channel linearly and directly, the idea of similarity matrix is utilized. More specifically, first identify feature matrices from each color channel as Fic1,Fic2,Fic3,where C1, C2, C3 represent each color channel. By collecting them as a larger matrix Fi=[Fic1,Fic2,Fic3], similarity matrix between Fi and Fj is applied .

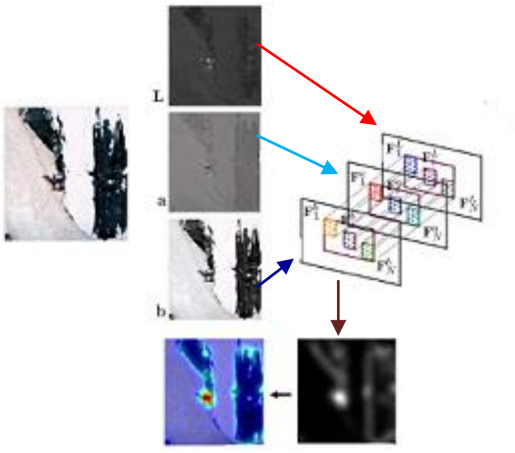

Fig 3. Saliency Detection in a color image

Then, the saliency map from color channels can be analogously defined by using Equation(3) as follows,

$$S_i = p(F \mid y_i = 1) = \frac{1}{\sum_{j=1}^{N} \exp\left(\frac{-1+\rho(F_i, F_j)}{\sigma^2}\right)} \qquad (3)$$

D. Calculation of Similarity Matrix

Use of similarity matrix provides consistent results regardless of color spaces and helps to avoid drawback of fusion methods. Saliency at a pixel Xi is measured using the conditional density of the feature matrix at that position: Si = P(F|yi = 1). Hence, the task at hand is to estimate P(F|yi = 1) over i = 1, · · · ,M. In general, the Parzen density estimator is a simple and generally accurate non-parametric density estimation method. However, in higher dimensions and with an expected long-tail dispersion, Parzen density estimator with an isotropic kernel is not the most appropriate method. The LSK features tend to generically come from long-tailed dispersions, and as such, there are generally no tight clusters in the feature space. While estimating a probability density at a particular feature point, for where L is the number of vectorized LSKs (f 's) employed in the feature matrix, the isotropic kernel centered on that feature point will spread its density mass equally along all the feature space directions, thus giving too much emphasis to irrelevant regions of space and too little along the manifold. Earlier studies [19] also pointed out this problem. This motivates to use a locally data-adaptive version of the kernel density estimator. Center+surrounding regions is used to compute similarity matrix which is considered as a local neighborhood. i.e., N << M. Red values in saliency map represent higher saliency, while blue values mean lower saliency. P(F|yi = 1) at Xi as a center value of a normalized adaptive kernel (weight function) G(·) computed in the center surrounding region as follows:

$$p(F \mid y_i = 1) = \frac{G_i(\overline{F_i} - \overline{F}_j)}{\sum_{j=1}^{N} G_i(\overline{F_i} - \overline{F_J})} \qquad (4)$$

The Equation(4) is used to overcome the disadvantages of the conventional Euclidean distance which is sensitive to outliers.

E. Generation of Saliency Map

Saliency is measured in terms of how much pixel it stands out from its surroundings. To formalize saliency at each pixel, let the binary random variable yi denote whether a pixel position Xi = [X1, X2]iT is salient or not as follows:

$$\text{Yi} = \begin{cases} 1, \text{if Xi is salient,} \\ 0, \text{otherwise,} \end{cases} \qquad (5)$$

where i = 1, · · · ,M, and M is the total number of pixels in the entire image. Saliency at pixel position Xi as a posterior probability Pr(yi = 1|F) as follows:

$$\text{Sr} = \text{Pr}(\text{yi} = 1|\text{F}) \qquad (6)$$

where the feature matrix, Fi = [fi1,…….,fiL] at pixel of interest Xi(center feature,) contains a set of feature vectors (fi) in a local neighborhood where L is the number of features in that neighborhood. In turn, the larger collection of features F = [F1,….,FN] is a matrix containing features not only from the center, but also a surrounding region (center+surround region,)N is the number of feature matrices in the center+ surround region. Using Bayes' theorem, Equation (6) can be written as

$$S_i = P_r(y_i = 1 \mid F) = \frac{p(F \mid y_i = 1) P_r(y_i = 1)}{p(F)} \qquad (7)$$

By assuming that 1) a-priori Pr(yi = 1), every pixel is considered to be equally likely to be salient; and 2) p(F) are uniform over features, the saliency we defined boils down to the conditional probability density p(F|yi = 1). The conditional probability density is estimated using p(F|yi = 1). Gao et al.[21]and Zhang [22]et al.have tried to fit a marginal density of local feature vectors p(f) to a generalized Gaussian distribution. In the proposed method we approximate the conditional density function p(F|yi = 1) based on free dispersion of kernel density estimation. While Itti and Baldi computed, as a measure of saliency, KL divergence between a prior and a posterior dispersion, In this proposed method we explicitly estimate the likelihood function using free parameter dispersion of kernel density estimation.

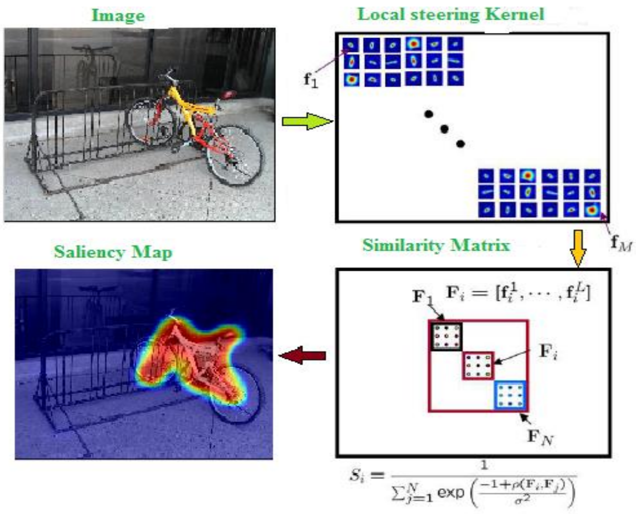
Image
Local steering Kernel
f_1
f_M
Saliency Map
Similarity Matrix
F_1
$F_i = [f_i^1, \cdots, f_i^L]$
F_i
F_N
$S_i = \frac{1}{\sum_{j=1}^{N} \exp\left(\frac{-1+\rho(F_i,F_j)}{\sigma^2}\right)}$

Fig.4. Graphical overview of Saliency Detection Process

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. Simulation Result Analysis

The proposed method evaluates spacetime saliency algorithm on the human fixation video data from Itti et al. This data set consists of a total of 520 human eye-tracking data traces recorded from 8 distinct subjects watching 50 different videos.Each video has a resolution of size 640X480. When comparing the proposed model against Bayesian Surprise and SUNDAY. Itti et al is also center-biased and Bayesian Surprise (Itti & Baldi) is corrupted by edge effects which resulted in relatively higher performance than it should have. For the evaluation of the algorithm, we first compute true positives from the saliency maps based on the human eye movement fixation points. In order to calculate false positives from the saliency maps, we use the human fixation points from frames of other videos by permuting the order of video. This permutation of images is repeated 10 times. Each time, we compute KL divergence between the histograms of true positives and false positives and average them over 10 trials. When it comes to calculating the area under the ROC curve, we compute detection rates and false alarm rates by thresholding histograms of true positives and false positives at each time of shuffling. The proposed model is simple, but very fast and powerful. In terms of time complexity, a typical run time takes about 8 minutes Zhang et.al reported that their method runs in Matlab on a video of about 500 frames in minutes on a Pentium 4, 3.8 GHz dual core PC with 1 GB RAM) on a video of size of 640X480 with about 500 frames while Bayesian Surprise requires hours because there are 432,000 dispersions that must be updated with each frame.

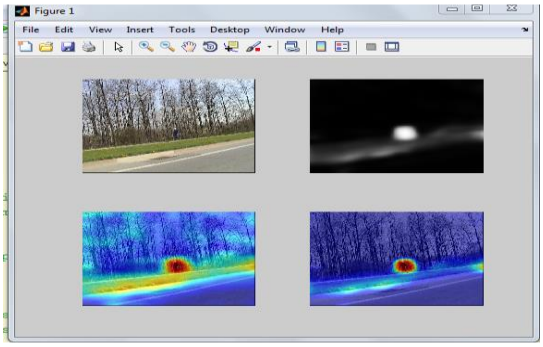

Fig 5. Saliency Map Generation of an image

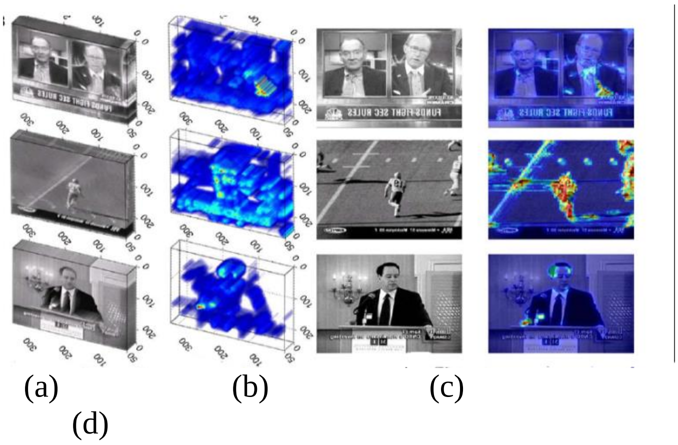

Fig 6. (a)Video Clip (b)Space-time Saliency map (c)Frames from the video clip (d) Saliency Map

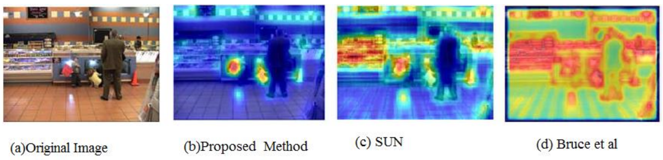

Fig 7. Comparative Results

IV. CONCLUSION AND FUTURE WORK

A bottom-up saliency detection algorithm is proposed by employing local steering kernels and by using a free dispersion of parameter kernel density estimation based on Similarity Matrix. The proposed method can automatically detect salient objects in the given image and in videos. The proposed method is practically attractive because it is parameter independent and robust to the uncertainty in data. Due to its resistance to noise and other systemic disruptions, the present model can be quite effective in other applications such as image quality assessment and video summarization.

REFERENCES

[1]R. Fergus, P. Perona, and A. Zisserman, “Object class recognition by unsupervised scale-invariant learning,” presented at the CVPR, 2003.

[2] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Shum, “Learning to detect a salient object,” presented at the CVPR, 2007.

[3] A. Treisman and G. Gelade, “A feature-integration theory of attention,” Cogn. Psych., vol. 12, no. 1, pp. 97–136, 1980

[4] J.Wolfe, “Guided search 2.0: A revised model of guided search,” Psychonomic Bull. Rev., vol. 1, no. 2, pp. 202–238, 1994

[5] C. Koch and S. Ullman, “Shifts in selection in visual attention: Toward the underlying neural circuitry,” Human Neurobiol., vol. 4, no. 4, pp. 219–227, 1985.

[6] L. Itti, C.Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[7] R. Rensink, “Seeing, sensing, and scrutinizing,” Vis. Res., vol. 40, no. 10-12, pp. 1469–1487, 2000.

[8] D.Walther and C. Koch, “Modeling attention to salient proto-objects,” Neural Netw., vol. 19, pp. 1395–1407, 2006.

[9] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch, “Attentional selection for object recognition–A gentle way,” Lecture Notes in Computer Science, vol. 2525, no. 1, pp. 472–479, 2002

[10] X. Hou and L. Zhang, “Saliency detection: A spectral residual approach,” presented at the CVPR, 2007

[11]D. Gao, V. Mahadevan, and N. Vasconcelos. “On the plausibility of the discriminant center-surround hypothesis for visual saliency. Journal of Vision, 8(7):13,1–18, 2008.

[12] N. D. Bruce and J. K. Tsotsos, “Saliency based on information maximization,” presented at the NIPS, 2005.

[13] D. Gao, V. Mahadevan, and N. Vasconcelos, “The discriminant center surround hypothesis for bottom-up saliency,” presented at the NIPS, 2007.

[14] J. Harel, C. Koch, and P. Perona, “Graph-based visual saliency,” presented at the NIPS, 2006.

[15]G. Backer, B. Mertsching, and M. Bollmann, “Data- and model-driven gaze control for an active-vision system,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 12, pp. 1415–1429, Dec. 2001.

[16] M. Z. Aziz and B. Mertsching, “Fast and robust generation of feature maps for region-based visual attention,” IEEE Trans. Image Process., vol. 17, no. 5, pp. 5, 633–644, May 2008.

[17] L. Itti, “Automatic foveation for video compression using a neurobiological model of visual attention,” IEEE Trans. Image Process., vol. 13, no. 10, pp. 1304–1318, Oct. 2004.

[18] A.Oliva, A. Torralba,M. Castelhano, and J. Henderson. “Topdown control of visual attention in object detection. In Proceedings of International Conference on Image Processing, pages 253–256, 2003.

[19] O. Meur, P. L. Callet, and D. Barba.” Predicting visual fixations on video based on low-level visual features. Vision on video based on low-level visual features”. Vision Research, 47:2483–2498, 2007