ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Salient Object Detection in Videos Based on SPATIO-Temporal Saliency Maps and Colour Features

U.Swamy Kumar
PG Scholar
Department of ECE,
K.S.R.M College of Engineering (Autonomous),
Kadapa.

Smt. G.Hemalatha, M.Tech
Associate Professor,
Department of ECE,
K.S.R.M College of Engineering (Autonomous),
Kadapa.

ABSTRACT

Salient object detection in videos is challenging because of the competing motion in the background, resulting from camera tracking an object of interest, or motion of objects in the foreground. The authors present a fast method to detect salient video objects using particle filters, which are guided by spatio-temporal saliency maps and color feature with the ability to quickly recover from false detections. The proposed method for generating spatial and motion saliency maps is based on comparing local features with dominant features present in the frame.

A region is marked salient if there is a large difference between local and dominant features. For spatial saliency, hue and saturation features are used, while for motion saliency, optical flow vectors are used as features. Experimental results on standard datasets for video segmentation and for saliency detection show superior performance over state-of-the-art methods.

INTRODUCTION

Modern day life has overwhelming amount visual data and information available and created every minute. This growth in image data has led to new challenges of processing them fast and extracting correct information, so as to facilitate different tasks from image search to image compression and transmission over network. One specie problem of computer vision algorithms used for extracting information from images, is to and objects of interest in an image.

Human visual system has an immense capability to extract important information from a scene. This ability enables humans to focus their limited perceptual and cognitive resources on the most pertinent subset of the available visual data, facilitating learning and survival in everyday life. This amazing ability is known as visual saliency (Itti et al. (1998)). Hence for a computer vision system, it is important to detect saliency so that the resources can be utilized properly to process important information. Applications range from object detection or Content Based Image Retrieval (CBIR), face or human re-identication and video tracking.

Motivation

Saliency is the ability or quality of a region in an image to standout (or be prominent) from the rest of the scene and grab our attention. Saliency can be either stimulus driven or task specific. The former one is known as bottom-up saliency while the later species top-down saliency and leads to visual search. Bottom-up saliency can be interpreted as a filter which allows only important visual information to grab the attention for further processing[1],[2].

In our work, we concentrate on bottom-up salient object detection. Saliency is a particularly useful concept when considering bottom-up feature extraction, since one must and what is significant in an image from the scene data alone. In such circumstances, the role of context becomes extremely important. That is to say that saliency can be described as a relative measure of importance. Hence, the bottom-up saliency can be interpreted as its state or quality of standing out (relative to other stimuli) in a scene.

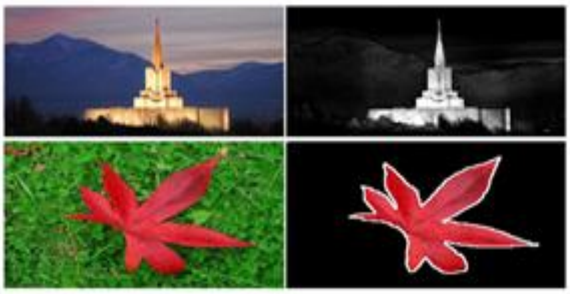

Fig: 1: saliency map generation

Figure 1 The top row shows an example of saliency map generated from the image (left) and the bottom row depicts an ideal segmentation of the object in the image (left).

As a result, a salient stimulus will often pop-out to the observer, such as a red dot in a yield of green dots, an oblique bar among a set of vertical bars, a dickering message indicator of an answering machine, or a fast moving object in a scene with mostly static or slow moving objects. An important direct effect of the saliency mechanism is that it helps the visual perceptual system to quickly filter and organize useful visual information, necessary for object recognition and/or scene understanding.

We propose two dierent methods for the task. First is based on low-level perceptual features. Second combines the low-level saliency with generic object specific cues in a graphical model based approach. Both the methods are thoroughly evaluated against state-of-the-art methods [1][3] (Cheng et al. (2011); Perazzi et al. (2012); Carreira and Sminchisescu (2010)) on challenging benchmark datasets and found to produce superior results.

Objective and Scope

The objective of the thesis is to device an efficient salient object detection method that can facilitate as a pre-processing step for many of the previously mentioned tasks. Further, the method must be unsupervised so that it can detect any generic object. Moreover, it has to be computationally efficient to ensure fast processing, considering the huge amount of available data. As already discussed bottom-up saliency can be characterized by the ability to pop-out in a scene. Consequently, most saliency detection methods in literature (Achanta et al. (2009); Goferman et al. (2010); Cheng et al. (2011); Perazzi et al. (2012); Li et al. (2013); Yang et al. (2013); Jiang et al. (2013)) propose a model by exploiting rarity of features. But, as also mentioned by Wei et al. (2012); Zhu et al. (2014) only feature rarity based approach is not enough to extract salient regions from natural images of varying scene conditions. We identify this shortcoming in the rarity of feature based approach and exploit boundary prior as a cue to implement our first method of saliency detection. Further, class independent object segmentation has recently gained importance in the Computer Vision community (Carreira and Sminchisescu (2010); Endres and Hoiem (2010)). In this context, Alexe et al. (2012) had addressed the problem of detecting generic objects and defined objectness properties as likelihood of a region belonging to an object. But it gives bounding boxes rather than pixel level segmentation output. However their precision is very low, as a lot of background regions are proposed as objects. Optimization is based on intersection-over-union criteria (Endres and Hoiem (2010)) to rank the maps, but the results show that the top-most map generally contains almost half of the image. Hence, we propose an algorithm to generate a single map segmenting only the objects of interest, using saliency and objectness on a conditional random field (CRF). The focus of this thesis is on one of the visual capabilities of human - finding objects of interest in images.

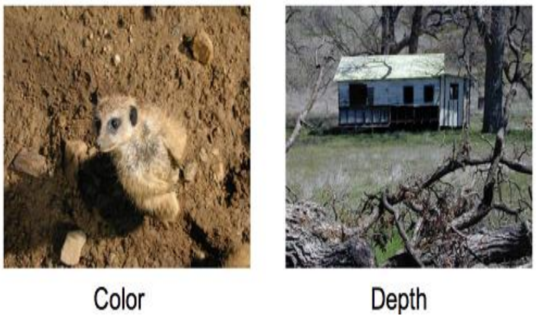

Fig: 2: saliency detection, illustrated using samples from saliency dataset

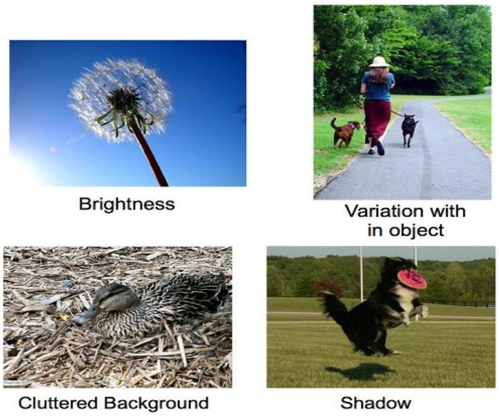

Fig: 3:Repeated Distractors in background or foreground

Figure 2: Dierent challenges in saliency detection, illustrated using samples from saliency dataset, MSRA B (Achanta et al. (2009)).

PROBLEM DEFINITION AND CHALLENGES

The problem we address in the thesis can be defined in short, as: Given a natural scene, detect one or more regions of interest (ROI) which contain the salient objects in a scene. The method must be unsupervised with no training sample for classes of objects available. Parameters of any optimization function may be learned using a part of another dataset, or verification subset of the same. Although the problem is similar to unsupervised foreground segmentation, it differs in the context of features which is mostly inspired by biological motivation. Some examples of finding objects of interest are presented in Figure 1.This is a challenging task, because objects of interest are detected without any prior knowledge about them purely based on unsupervised stimulus driven perceptual cues. Single features such as, color, brightness, depth alone is not helpful to solve the problem. It suffers from all the challenges that any computer vision problem faces and are illustrated in Figure2.Further, when finding saliency for challenging datasets like PASCAL to facilitate later process of object detection or recognition, segmenting the object of interest becomes even harder. A thorough study of different method and samples from the dataset reveals the following challenges, apart from the factors already depicted in Figure 2:

1. Only a small part of an object is present on the boundary of an image;
2. Objects with large holes, such as cycle wheel;
3. Repeated Distractors in background or foreground.

These are illustrated with respective samples in Figure 3

Contribution

The central contribution of the thesis is pixel accurate localization of the object of interest. The saliency map provided by our proposed methods assign each pixel a saliency value in the range of 0 to 1, depicting their probability of being salient. Hence, it can be easily segmented by simple thresholding mechanism, to obtain the important or salient object. In the work described here, saliency is defined firstly in terms of spatial rarity of image feature, mainly color. Secondly, objectness is used in a graphical model for salient object segmentation. This can change the conventional way of extracting features from the whole image or searching objects in huge 4-dimensional (position, scale and aspect ratio) sliding window search space. It would help simulate the same logistics as human vision and improve both speed and accuracy of the computer vision tasks. Moreover, since we produce a probability value of each pixel being salient, the saliency map can also be utilized for identifying most salient regions for different tasks, for example placing an advertisement in a video. In the following sub-sections, we describe the methods proposed in the thesis in brief.

SALIENCY DETECTION BASED ON LOW-LEVEL PERCEPTUAL CUES

In the first formulation, we formulate three saliency criteria, namely: (i) graph based spectral rarity, (ii) spatial compactness and (iii) background prior. Then, a weighted combination of these three measures of saliency, produce a saliency map. A saliency map is represented as a gray scale image, where each pixel is assigned its probability of being salient.The first two terms named above are based on rarity of feature and the third term correspond to boundary prior. The idea of boundary prior is that the boundary of an image mostly contains background image elements or superpixels and background superpixels are spatially connected among themselves, but not with foreground ones. Graph-based spectral rarity assigns saliency based on rarity or uniqueness of a superpixel. This measure utilizes the spectral features (as defined by Ng et al. (2001)) using Laplacian of the superpixel graph.

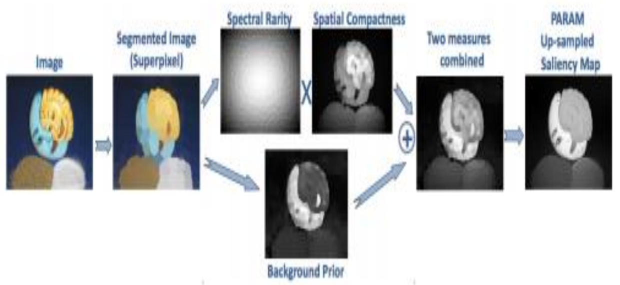

Fig: 4: Illustration of the sequence of stages of our proposed algorithm (PARAM) for saliency estimation, with an example from MSRA-B Dataset (Achanta et al. (2009)).

On the other hand, spatial compactness takes into account that a color belonging to a salient object would be grouped at a spatial location and thus the spatial variance of the color would be low. Whereas, background colors are generally distributed over the whole image and score low on spatial compactness.

Our formulation models background prior using a Gaussian Mixture Model (GMM). All the superpixels touching the boundary of an image are modeled by GMM in Lab color space. Saliency of a superpixel is measured as the sum of the distances from the GMM modes weighted by the particular mixture coefficient. Since, most of the boundary superpixel would be background, big GMM modes with high value of mixture coefficient belongs to background colors and thus the mentioned distance gives a good measure of saliency. A non-linear weighted combination of these three different cues is used to compute the final saliency map. Also, binary segmentation maps are generated for quantitative evaluation of performance using an adaptive threshold. An illustration of the complete flow chart of this proposed saliency detection method, which is named as PARAM (background Prior and Rarity for saliency Modeling), is depicted in Figure .4.

SALIENT OBJECT SEGMENTATION IN NATURAL IMAGES

Next we propose a Salient Object Segmentation method that captures the same visual processing hierarchy as in the human visual system. Our goal is to localize objects independent of its category by formulating an unsupervised algorithm.

Recent saliency detection methods show high performance in saliency datasets, but they fail to perform well when tested using natural image datasets like PASCAL (Everingham et al. (2012)). There are two reasons behind this. First, these methods use only low-level perceptual cues such as, center surround operations (Itti et al. (1998)), local and global contrast (Goferman et al. (2010); Cheng et al. (2011)), uniqueness and color distribution (Perazzi et al. (2012)) and boundary prior (Yang et al. (2013); Wei et al. (2012)). Second, there is typically a huge dataset bias which ensures the presence of only a single object at the center of an image. Moreover, in saliency datasets the objects are in high contrast with respect to the background. Hence, this class of methods does not scale up for more natural images such as in the case of PASCAL segmentation dataset (Everingham et al. (2012)).

Here, we exploit the saliency (PARAM) described in the previous section along with objectness cues. We formulate two simple but effective objectness factors: geometric constraint and distribution of edges in the image. These two features are respectively modeled as boundedness and edge-density. To compute these factors we exploit the edge map produced by Dollar and Zitnick (2013). They take a structured learning based prediction on random forest to produce a high-quality edge probability map. Since they do a direct inferencing, the method is computationally efficient. Boundedness captures the extent to which a superpixel is bounded by strong edges on all four directions. Whereas, edge-density computes how much cluttered or smooth a particular superpixel is. As described by Alexe et al. (2012), very smooth or highly cluttered regions generally belong to the background.

Next, a graphical model based approach is used for proper spatial propagation of these priors. The image cue priors form the unary term and we formulate a sub modular edge cost or pairwise term to specify the CRF (Laerty et al. (2001)). Hence, an exact inference is done efficiently using graph-cut. We employ a margin rescaled algorithm, as explained by Szummer et al. (2008), to learn the CRF parameters. It is a max-margin structured learning based approach. The benefit of the formulation is that, it takes into account how far a predicted label is from its ground truth, and the margin is adapted based on how much competing labeling differ from ground truth. Results of the proposed method shows superior performance on PASCAL segmentation dataset (Everingham et al. (2012)) when compared against many recent methods.

SALIENT OBJECT DETECTION

We employ a particle filter for its ability to approximate the posterior distribution of a system based on a finite set of weighted samples. Particle filters are also highly robust to partial occlusion and computationally inexpensive. The weight of each particle is initialized using a uniform distribution. The first set of particles is initialized around the centre of the first frame of the video. Weights are subsequently calculated as the weighted sum of distance measures of the candidate regions to the reference distribution. The spatio-temporal saliency and the color maps are used to calculate the weight of the samples, which allows subsequent iterations to move the particles closer to the most salient object.

In the proposed framework, we detect only one object of interest. Fig. 1 illustrates a workflow of how the particle filter framework is used to detect the salient object. Color versions of all the Figures used in this paper are available online.

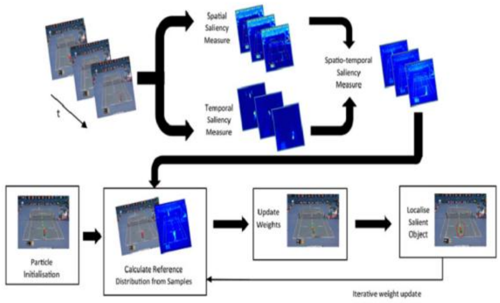

Fig: 5 Illustration of the work flow of the proposed framework

MOTION SALIENCY MAP

Studies have shown that motion plays a major role in garnering attention and to outweigh low-level features such as orientation, intensity and texture in videos. Yantis et al. have shown that the HVS is particularly sensitive to isolated abrupt stimulus and relative movement of objects which refers to the contrast in motion within a spatial neighborhood. In a similar spirit, we identify salient motion from the contrast or the relative movement of objects in the scene calculated as

$$\mathrm{SMot}(m_k) = \sum_{\forall m \in M} p(m) ||m_k - m||_1 \qquad (10)$$

Where M is the set of dominant motion magnitudes computed from optical flow vector velocities that are obtained according to and mk is the motion magnitude at pixel k.

SPATIAL SALIENCY MAP

Spatial attention cues influence the attention process when similar motion dominates the frame, for example, swaying of trees in the background or when the foreground objects have similar motion, for example, a crowd walking in the same direction. The spatial saliency map of a frame is calculated based on the colour contrast according to the formulation in (9). Unlike motion, colour contrast is influenced by the compactness of similar colour pixels [50]. We adopt the method proposed by Morse et al. [51] to calculate the dominant colours using hue and saturation maps from the HSV image. The effect of the saturation values are included in the hue histogram which is hence defined as H(b) = (x,y)[Ib S(x, y), where H(b) is the histogram value in the bth bin, S(x, y) is the saturation value at location (x, y) and Ib is the set of image coordinates having hues corresponding to the b_{th} bin. The dominant hues are calculated in a similar manner as dominant motion by obtaining the local maxima values from a smoothed histogram H(b). The spatial saliency measure of a pixel k is

$$\mathrm{SSp}(c_k) = \sum_{\forall c\in C} p(c)||c_k - c||_1 \qquad (11)$$

where C is the set of dominant hues and ck is the hue value at pixel k. Fig. .2 shows frame from three different videos and their corresponding spatial, motion and spatio-temporal saliency maps. The top row shows a frame from a video of a sauntering rhinoceros, shot with a stationary camera while the middle and bottom rows are videos of a wolf roaming in the forest and a skier performing a stunt, respectively. The last two videos are of tracking shots

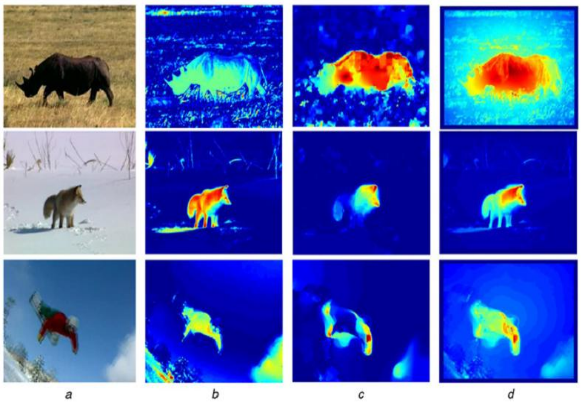

Fig: 6: Saliency maps a Original frame b Spatial saliency map c Motion saliency map d Spatio-temporal saliency map

In the top row, the movement of the rhinoceros is large when compared to the background flutter in the grass. Hence, the animal’s motion is assigned a larger saliency than that of the swaying grass in the background. In the video with the wolf (middle row), the pixels present in the background have large motion primarily owing to the camera tracking the wolf while the motion magnitudes of the pixels on the wolf are low. The motion saliency map assigns a high confidence to the motion of the wolf owing to the large difference in the magnitudes between the tracked wolf and the dominant background motion. The frame shown in the bottom row of Fig. 4 is from a video of a camera tracking a skier performing a stunt. Similar to the case above, the skier is executing a somersault when his hand moves faster than the rest of his body, generating a large saliency measure along his arm. Thus, it can be seen that the proposed motion saliency measure is able to successfully identify the salient motion present from videos shot with a stationary camera or a tracking camera.

SPATIO TEMPORAL SALIENCY MAP

The spatio-temporal saliency map combines the spatial and motion saliency maps calculated for each frame in such a way that the motion map gets a larger weight if there is high motion contrast in the sequence while the spatial saliency map gets a larger weight if the motion contrast is low. This is formulated as

$$\mathrm{STSal}(I) = \alpha \times \mathrm{SMot}(I) + (1-\alpha) \times \mathrm{SSp}(I) \qquad (12)$$

Where α is an adaptive weight given by

$$\alpha = \frac{\mathrm{median}(\mathrm{SMot}(I))}{\max(\mathrm{SMot}(I))} \qquad (13)$$

If a large number of pixels have high motion saliency, then the median is closer to the maximum. On the other hand, if there are fewer pixels the at are closer to the maximum motion saliency measure, the median returns a lower value and α evaluates to a lower value indicating that the influence of motion is not large enough for the motion saliency to dominate the attention process The spatio-temporal saliency map generated for the three video sequences discussed earlier are shown in Fig. 4 d. The motion adaptive weights allow for a better estimate of the spatio-temporal saliency measure when the video is influenced by a large motion contrast. The top row in Fig. 4 d is an example where the motion cues play a major role in the spatio-temporal saliency measure as the dominant hues are not very different for the foreground salient object and background. The frames shown in the middle and bottom row of Fig. 4 d are

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

examples where the camera tracks an object of interest. In this case, the dominant motion pixels are present in the background although the motion saliency is not large enough until the object had its own local motion. In these two cases, the spatial saliency map takes over the majority of the spatio-temporal saliency measure as the motion contrast is not high enough for the motion map to get control

SIMULATION RESULTS

The step by step procedure for simulation results is

- Initially took a small video (sample video).
- The sample video is divided into the number of frames based on the motion saliency map.

Fig: 7: sample frames from sample videos

- Then store all frames in to a same folder
- This will be the input.
- In output we get the motion able parts as a dot points. (particles)

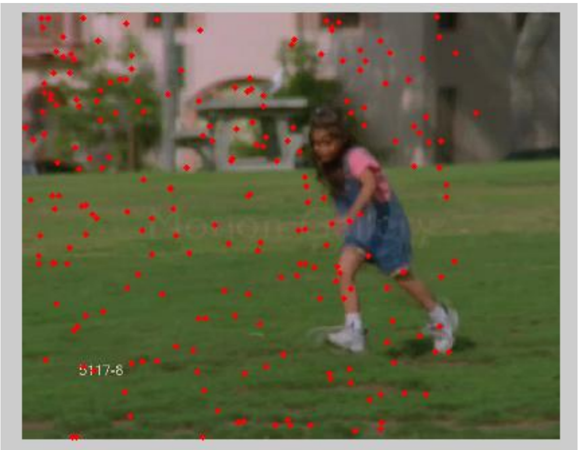

Fig: 8: showing particles.

CONCLUSION

We have presented an algorithm for salient object detection in videos based on particle filters that uses spatio-temporal saliency maps and colour as cues. The performance is evaluated on segmentation datasets. We also develop a simple algorithm to generate spatio-temporal saliency map that outperforms many state-of-the-art methods. As a future work, we extend the results of the salient object detection framework to intelligently resize frames of a video.

REFERENCES

[1] Han, S.-H., Jung, G.-D., Lee, S.-Y., Hong, Y.-P., Lee, S.-H.: ‘Automatic salient object segmentation using saliency map and color segmentation’, J. Central South Univ., 2013, 20, (9), pp. 2407 –2413

[2] Itti, L., Koch, C., Niebur, E.: ‘A model of saliency-based visual attention for rapid scene analyses, IEEE Trans Pattern Anal. Mach. Intell., 1998, 20, (11), pp. 1254 –1259

[3] Simakov, D., Caspi, Y., Shechtman, E., Irani, M.: ‘Summarizing visual data using bidirectional similarity ’. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 2008

[4] Hou, X., Zhang, L.: ‘Saliency detection: a spectral residual approach’. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 2007

[5] Duncan, K., Sarkar, S.: ‘Saliency in images and video: a brief survey’, IET Comput. Vis., 2012, 6, (6), pp. 514 –523

[6] Itti, L., Baldi, P.: ‘A principled approach to detecting surprising events in video ’. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, San Diego, CA, USA, 2005, pp. 631 –637

[7] Mahadevan, V., Vasconcelos, N.: ‘Spatiotemporal saliency in dynamic scenes’, IEEE Trans. Pattern Anal. Mach. Intell., 32, (1), pp. 171 –177

[8] Gopalakrishnan, V., Hu, Y., Rajan, D: ‘Sustained observability for salient motion detection’. Proc. 10th Asian Conf. on Computer Vision, Queenstown, New Zealand, 2010, pp. 732 –743

[9] Muthuswamy, K., Rajan, D.: ‘Salient motion detection through state controllability ’. Proc. IEEE Int. Conf on Acoustics, Speech and Signal Processing, Kyoto, Japan, 2012, pp. 1465 –1468

[10] Xia, Y., Hu, R., Wang, Z.: ‘Salient map extraction based on motion history map ’. Proc. 4th Int, Congress on Image and Signal Processing, Shanghai, China, 2011, pp. 427 –430.