International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Efficient Salient Region Detection: A Review on a Data Driven Distinctiveness for Salient Region Detection in Natural Images

[1]Aarzoo Patwa, [2]Dr. Ashish Kumar Khare

[1,2]Computer Science and Engineering, LNCTS, BHOPAL, INDIA

ABSTRACT—

More comprehensive information, such as depth cues, inter-image correspondence, or spatiotemporal relationships, is now available on image saliency detection models that focus on extracting salient objects from images by combining colour and depth information, thanks to advancements in acquisition technology. Saliency is a crucial component of image information that can easily limit one's attention. The goal of saliency detection is to effectively highlight the important things while suppressing the irrelevant ones. Saliency detection is a dynamic study domain in the science of image processing that recognizes and restricts the part of a picture that is salient to the Human Visual System (HVS). Over the last few decades, computational modeling of visual attention has been the latest research topic. The framework of modern computer vision systems requires the detection of salient regions.

Index Terms— Salient regions, Saliency Detection, Feature Extraction

I. Introduction

Various strategies have been developed over time to predict visually prominent information in photos that can also keep in contact with particularly revealing portions of the image. These salient regions have traditionally been measured as local phenomena, with the salient regions acting as local extrema in relation to their immediate neighbours. A number of computational strategies in computer vision are planned based on primates' visual attention. These visual saliency models have demonstrated to be useful in a variety of applications, including robot localization [1], salient object recognition [2], object tracking [3], video compression [4], thumbnail generation [5,] and so on. [6] contains a full overview of the topic under consideration. In mandrill intelligence, there are two distinct brain pathways for fundamental visual attention. The objective-driven top-down [6] method is slow and relies on learning, knowledge, and recall. The sensor-driven bottom-up [6] approach, on the other hand, is rapid and only works with readily available motivation. Many computer vision algorithms employ a bottom-up approach to identifying important features in a group of images. Using this strategy would demand well-categorized encoding of several representation unpredictable's that might characterize image properties.

On the other hand, recent advances in computer vision literature have indicated that these important qualities can be considered in the context of the entire image (or data collection) and can be gathered using distinct methods. However, such an inference is not always possible due to the high-dimensional natural world of image properties.

Figure-1: Illustration of saliency detection.

Saliency is the state or attribute of an object that allows it to stand out from its surroundings [7]. By initially detecting the numerous items present in the image and then inferring whether they might stimulate visual attention, a computer system can be trained to determine the most visually salient regions in a top-down way. Alternatively, saliency can be conditioned in a bottom-up way that is totally focused on the items in the image. In brief, the feature they extract from photos is the level of detail (See Figure 2). Separate prototypes that are least similar other models in the image are removed using the saliency detection algorithm given. Low level features, which indicate local saliencies, are more comprehensive than higher level traits.

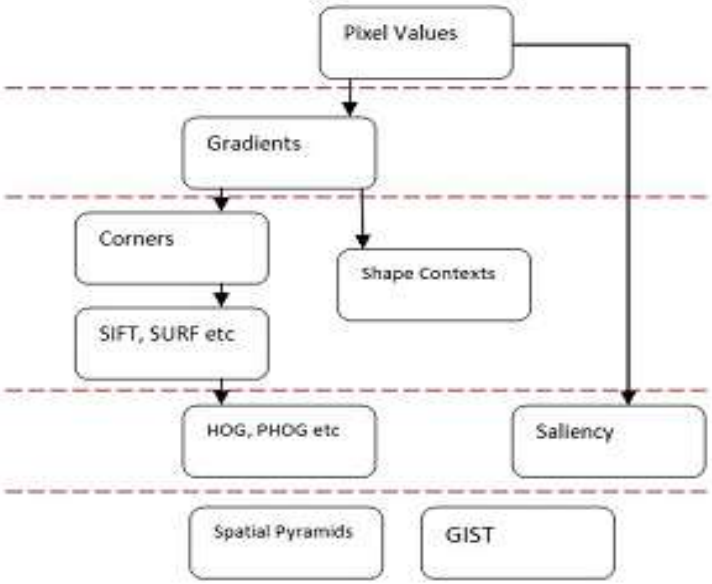


Figure-2: Hierarchy of image features.

There are three types of saliency approaches: solely bottom-up, top-down, and hybrid approaches that combine top-down and bottom-up modes [8]. Top-down approaches use a-priori known knowledge about the appearance of target objects, tasks, and high-level signals to drive saliency towards key image regions, whereas bottom-up approaches use image-based conspicuities to compute saliency.

II. Theoretical Concept

For well-organized observation, selecting only a subset of the available sensory information before more detailed processing is necessary. In the visual realm, this is often accomplished by suppressing information beyond the "centre of attention" zone. Because the specifics of the underlying biological visual system have yet to be revealed, the definition of the centre of attention remains opaque. However, using a serialised analysis method, primate brains can attain great performance in scene interpretation and analysis with limited processing resources. Visual spots that "pop-out" or are conspicuous in comparison to the rest of the scene receive attention at the top of this sequential process, in an unconscious manner. This trait is extremely important in monkeys' behaviour since it helps them to detect unexpected predators almost instantaneously. Visual processing complexity can be reduced greatly by filtering out extraneous and duplicated visual information and determining the most relevant components of an image. Adaptive image compression, object detection and recognition [9], thumbnail generation [10], content-aware image re-targeting [11], photo collage [12], and non-photorealistic rendering are just a few of the important computer vision and realistic applications that could benefit from an effective computational model for automatically generating saliency maps from representations. Computational saliency modeling research can potentially reveal computational features of the underlying neurological attention mechanisms.

III. The Meaning of Saliency

The recognition of salient areas in a picture is one of the most difficult problems in the field of computer vision. In the absence of any external supervision, psychovisual tests [13] reveal that attention is drawn to visually salient places in the image. These principles define visual coherence from a variety of perspectives. Visual saliency, often known as saliency, is concerned with finding fascinating spots that a human observer would focus on during a first glance. The points of interest are commonly referred to as "salient" points. The word "salient" means "prominent," "conspicuous," and "projecting or pointing outward," according to Webster's Dictionary. It is also something that stands out from the rest of the environment. The above-mentioned detectors' interest points are not necessarily conspicuous in this intuitive sense. The importance of a feature is determined by its context, which these feature detectors do not take into account. The term "visual saliency" usually refers to a quality of a "point" in an image (picture) that allows it to be absorbed. Most visual saliency detection representations are driven by the human visual system, and they tend to recreate energetic changes in brain connections for scene perception saliency by claiming that not all interest locations are salient in the intuitive sense. The context, i.e. the rest of the image, strongly influences whether or not a point or a patch in an image is salient. As a result, truly noteworthy points or patches may not be repeatable, and so may not be ideal interest points.

A. SALIENCY APPROACH

- **What is saliency?**

The work in this research is centered on the principle of extracting as much information from a scene as possible before using top-down discriminative models. I've already discussed the drawbacks of using models, especially in natural settings. As a result, the work's aim is to maximize knowledge extraction from video data by determining its underlying spatial-temporal dynamics. The primary idea of the research is to use a saliency measure to determine what is significant in a geographical, temporal, and/or spatial-temporal context. If a mechanism for quantifying saliency is available, the information or data from a certain scene has been translated to a feature domain in which some amount of inference about the raw data has been extracted. When examining bottom-up feature extraction, saliency is particularly beneficial because it is necessary to determine what is relevant in an image or sequence based solely on scene data. The importance of context becomes much more significant in these situations. That is, saliency can only be represented as a measure of relative importance.

Saliency is first characterized in terms of spatial, then spatio-temporal homogeneity, in the work detailed below. As a result, zones of spatial or spatio-temporal homogeneity must unavoidably come to an end or encounter a contour where the pixel intensities differ more than expected. Edge detection provides us with a spatial and/or temporal framework in which to analyse the region. At this point, it's crucial to take a brief detour to talk about the importance of edges in terms of visual saliency. For locating relevant features within a scene, early feature extraction approaches tended to focus on edge extraction methods. This is an excellent approach of determining spatial structure in simple circumstances. In the end, we're more interested in figuring out what the object is that the edges come from. For non-rigid objects, constructing such information from edges is particularly difficult. Finding edges in order to uncover spatially homogeneous things appears to be an overuse of computer resources, given that the natural world is mostly made up of deformable shapes.

- **Binding features by co-occurrence**

Finding an efficient means of tying characteristics together is crucial when understanding photos or videos. Combining features geographically in a meaningful way is a difficult challenge in and of itself; one can only image how difficult it would be to do it temporally. Furthermore, if we are interested in modelling activities, it quickly becomes clear that the same activities can take place at quite different spatial separations. As a result, location-invariant approaches must be used to represent physically isolated but temporally connected activity. To this purpose, co-occurrence is a useful tool for tying together characteristics that aren't always geographically close together or distinguished by a distinct topographical layout. Another benefit of co-occurrence across temporally associated activities is that no explicit assumption of temporal ordering or synchrony is required. There is a lot of human-human contact in natural environments, especially in surveillance video. This varies a lot based on the scene topology and each person's gaze direction. Finding some loose temporal synchronization is a realistic technique to simulate the often observed cause-effect phenomena that is evident in interaction behavior when we are attempting to determine whether they are engaging with each other. That is, whether two persons are interacting or engaging with each other is not dependent on the sequence in which the co-occurring traits appeared.

- **Finding interactions and group behavior**

Finding interactions or noteworthy group behaviour is a difficult but fascinating problem. The majority of human activity recognition research has focused on individual activity recognition. Multi-person activity, on the other hand, is frequently of greater interest in natural surveillance scenarios since someone is more likely to be in immediate and preventable danger from their surroundings.

B. BOTTOM-UP SALIENT SPATIAL FEATURE SELECTION

Saliency is a critical feature of image or video comprehension, as explained in the previous article. When we look at the disparities between the human visual system and the most commonly used feature selection approaches, this becomes clear. We can approach the problem by looking at how most natural sceneries are congested. In such scenarios, the human visual system is limited to selecting only those aspects of a scene that we judge important enough to spend additional computing resources on. Such pre-attentive mechanisms are a natural and rational way of focusing computational load in order to comprehend spatial or temporal dynamics of a stimulus more quickly [7].

- **Saliency Map**

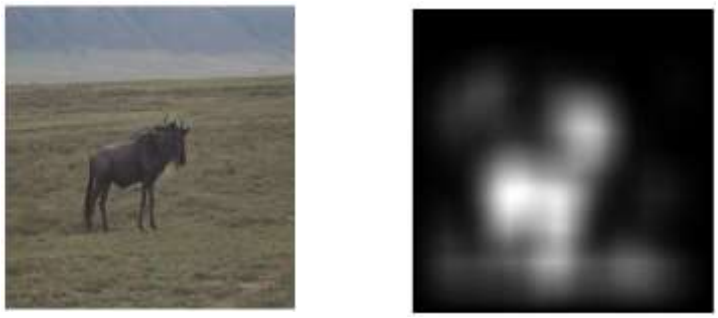
By focusing on low-level visual properties such as texture, a new measure of Visual Saliency is known in this technique. What's the big deal about texture information becoming aware of visual saliency? The answer is that texture provides crucial information about the "behaviour" of an image. According to the strategy provided here, the base for extracting salient regions is to emphasise texture rare events. The spatial distribution of important spots within a picture is investigated in order to show textural differences throughout the image. The levels of irregularity in both excellent and common areas can be quite high; in an excellent region, they will encounter a greater number of key points than in a common region, so key point density is used to distinguish between different texture occurrences and to identify the most important regions. As a result, we arrive at the crux of the problem: saliency can only be represented as a relative measure. Many various approaches have been tried to find a measure that accurately represents how 'prominent' some sections of the scene are.

For a given k, the saliency map SM is constructed as the absolute difference between the SDM values and the map's most frequent value MV, as described in their technique [14]. A SIFT Density Map (SDM) is a visual representation of the density of significant locations in an image that can reveal important details about the texture's consistency. A SIFT Density Map SDM (k) is created by counting the number of key points that characterise the scale of observation into a sliding window of size k x k.

$$SM(k) = |SDM(k) - MV(SDM(k))| \quad (2.1)$$

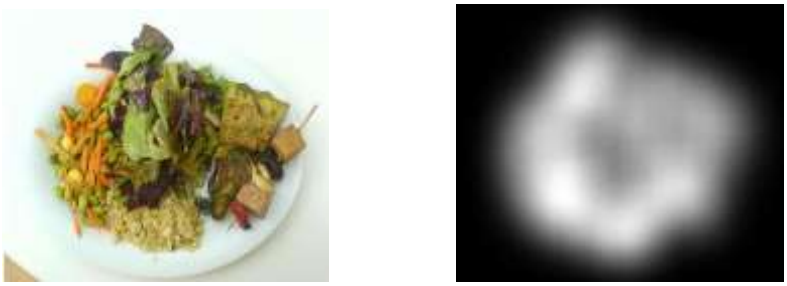
which is normalised in relation to the maximum value in order to limit SM values to [0,1].

The bulk of important places in the image are those that are conveyed to the SDM values with the greatest departure from the most common value, which are typically the image's most uncommon texture occurrences. This determines both the scenario where a textured object is the salient region, since it is surrounded by consistent areas (i.e. the most frequent value close to 0), and the case where a homogeneous region is surrounded by textured sections (i.e. the top most frequent value). (Figures 3 and 4).



(a) (b)

Figure-3: (a) homogeneous subject in a textured scene and (b) the corresponding Saliency Map



(a) (b)

Figure-4: (a) textured object in a homogeneous background and (b) the corresponding Saliency Map

- **Interest point detection**

In the 1960s, it was claimed that the biological visual system picks out basic elements including orientation information and scales. [15] also used consistency as a metric of saliency in their study on space-time interest points. This approach chose interest sites by looking at how often motion turning points were repeated throughout time. All of these strategies have the drawback of being able to pick out interest points reasonably reliably, but the points are only of interest in a local sense. It is impossible to select locations based on more global conceptions of saliency without a measure of how important the region is. Furthermore, they do not address the issue of how to combine the features into more unified objects. This is especially true because interest points are edges or turning points in space or time that can only ever represent a portion of an entity. When there are several items in the picture, each with an undetermined amount of feature points, grouping them together becomes a difficult task, especially when they are near together or overlap geographically. Furthermore, regardless of the global context in which the points are selected, these places of interest are selected using a binary procedure, which tends to lead to either over or under selection.

It would be far more convenient if conspicuous homogenous regions of interest could be extracted in one step, along with a trustworthy assessment of how important or prominent the region is in comparison to others. Furthermore, most of the methods discussed above ignore how the temporal evolution of interest points may cause some to become less salient than others as a result of their frequency of repeat. Over-selection could become a problem if more effective feature selection is not available in spatially cluttered image sequences, making these approaches intractable for interpreting cluttered outdoor scenes with dynamic background lighting changes.

- **Spatial saliency detection**

The mechanism by which saliency is considered is the major difference between interest point detection and spatial saliency detection. While interest point detectors attempt to choose local changes in the spatial data, saliency detectors select quantifiably discriminative regions of interest to perform a semantically higher degree of extraction of what is essentially relevant in a scene. That is, the features can be extracted in such a way that their representation is more closely tied to a symbolic or semantic meaning that can be associated with common items. It's more like the human visual system's concept of visual 'pop out,' which evaluates how characteristics might be joined together to generate a higher meaning than the individual parts. That is, saliency detection attempts to extract what is not just locally but also globally noteworthy. Local context is an essential element of the feature extraction

process, therefore determining the scale of salient regions based on a statistical measure of the data has an advantage. Because the entropy metric does not clearly represent one color/intensity distribution, the level of spatial homogeneity can be specified relative to a local environment. This is especially important since, regardless of how textured or varied the local spatial structure is, we would definitely want to select regions of interest that may equal to a greater semantic value.

IV. Literature Survey

The author of this research [16] must integrate the two methodologies and offer a strategy that utilises both sparse and nonlinear characteristic representation. They use independent component analysis (ICA) and covariant matrices to arrive at this conclusion. They use a geographically suitable centre surround difference (CSD) method to calculate saliency here. In the natural world, our sparse face appearance is adaptive; the ICA basis functions are learned to some extent at each image exhibition rather than being predetermined. They show how Adaptive Sparse Features, when combined with a CSD mechanism, might defer enhanced outcomes to permanent sparse demonstrations. They also provide you an idea about covariant matrices, which are made up of nonlinear colour integration alone are enough to proficiently approximation saliency from an image. On known datasets, the suggested twofold illustration approach is estimated alongside human eye fixation forecast response to psychological prototypes and salient object recognition. They conclude that having two types of demonstrations that complement each other results in improved saliency detection.

In this research [17], author propose Deep 3D Video Saliency, a new stereoscopic video saliency detection approach based on 3D convolution neural networks (Deep3DSaliency). Spatiotemporal Saliency Model (STSM) and Stereoscopic Saliency Aware Model (SSMAM) are two sub-models in the proposed network (SSAM). STSM uses three consecutive video frames as input to extract visual spatiotemporal information, while SSAM uses STSM's common parameters to infer depth and semantic features from the left and right video frames. An alternating optimization approach is used to learn the visual spatiotemporal characteristics from STSM and the depth and semantic information from SSAM. Finally, using 3D de-convolution, all of these saliency-related features are integrated for ultimate stereoscopic saliency detection. The suggested model outperforms other current models in saliency estimation for 3D video sequences, according to experimental results.

An MCDO approach for stereoscopic image saliency detection was proposed in this study [18]. Using an existing stereoscopic or 2D image saliency detection system, here they first create an initial saliency map. The original saliency map is then optimised using MCDO. In MCDO, a fully connected CRF is used to combine many cues such as colour, spatial position, and depth. The proposed optimization approach can be used with a variety of stereoscopic image saliency detection systems. We used eight saliency detection evaluation measures to assess the improvements. Experiments on three stereoscopic image saliency detection datasets show that MCDO enhances the state-of-the-art saliency detection performance for stereoscopic pictures by working for both stereoscopic and 2D saliency detection techniques. More signals may be added in the future to improve the efficacy of stereoscopic image saliency detection.

A DCT-domain-based contrast enhancement approach is proposed in this study [19]. This method uses local adaptive processing to greatly improve image contrast and colour information while maintaining acceptable image enhancement quality. The The display of a colour image is primarily determined by three factors: (i) brightness, (ii) contrast, and (iii) the original colour composition. While devising a traditional method for computing, the algorithm uses the DSR method to enhance a dark or low contrast image. Numerous earlier works on many algorithms have considered either the brightness, i.e. dynamic range correction, or the contrast, i.e. image sharpening procedures, or a mix of both qualities in some circumstances. However, there were few suggestions for preserving colours in the improved image. When compared to existing enhancement methods such as adaptive histogram equalisation, gamma correction, single-scale retinex, multi-scale retinex, modified high-pass filtering, multi-contrast enhancement, multi-contrast enhancement with dynamic range compression method, colour enhancement by scaling method, and multi-contrast enhancement with dynamic range compression method, colour enhancement by scaling method, The proposed technique delivers significant presentation in terms of comparative contrast enhancement, colorfulness, and visual perception of boosted image using edge-preserving multi-scale decomposition and common controls of well-liked imaging gadget. As a result, it can be concluded that the proposed DCT-based DSR technique outperforms existing image enhancing techniques in terms of contrast enhancement, visual information enhancement, and colour enhancement.

A dynamic stochastic resonance (DSR)-based technique for contrast enhancement of dark and low contrast pictures in the discrete wavelet transform (DWT) domain has been suggested in this study [20]. The conventional technique presents a stochastic resonance (SR)-based method that is improved by external noise accumulation and adaptive processing to achieve objective optimal concert. On the other hand, the proposed approach dynamic stochastic resonance within reach of an image's internal noise has been used for contrast enhancement. To depict the conditions of a wavelet coefficient as the movement of a particle in the double well, an equivalence of a dark picture domain to bistable double-well is applied. The humiliation caused by bad lighting is satisfied as noise, which is then employed to generate a noise-induced alteration of the image from low-contrast to high-contrast. In an iterative process, stochastic resonance is created in the approximation and detail coefficients, resulting in an increase in the variance and mean of the coefficient distribution. The selection of the best bistable system ensures optimal output response parameters. The signal-to-noise ratio of a traditional SR system is used to make parameter selection. Iteration is adaptively completed on achieving the desired perceptual quality, contrast quality, and colorfulness values. The suggested technique is evaluated using a variety of SR-based and non-SR-based methodologies, which indicates the method's potential and impressive presentation in terms of contrast quality, colour enhancement aspect, and visual information. These advancements can be aided by appropriate adjustments for application to bright images, and they can be furthered by adaptive region selection for processing.

The use of stochastic fluctuation, or noise, for image enhancement was examined in this study [20]. This method's distinguishing feature is the use of internal noise instead of external noise, as well as adaptive processing to achieve objective optimal performance. At the lowest iteration count, an iterative technique is used to achieve object value of routine metrics such as comparative contrast enhancement feature (F), perceptual quality measures (PQM),

and colour enhancement feature (CEF). When compared to existing SR-based and non-SR-based enhancement strategies in both the spatial and frequency domains, the suggested method is found to provide exceptional harmony in terms of contrast augmentation, perceptual superiority, and colorfulness.

Conclusion

In this study, we looked at a variety of saliency-based image algorithms to see how far they've come and how much better they've gotten. However, there are still a number of unresolved issues that need to be studied further. From natural images, the Human Visual System's (HVS) capacity to detect an object in an image is incredibly rapid and reliable, but how can a machine vision system find the salient regions? Several models and methods have been developed to handle this problem by extracting features in either the spatial or spectral domain with this data and executing various processes for improved features, noise reduction, and so on, resulting in improved visual quality. Because these procedures were proposed based on intuition and ideas derived from psychophysical studies, they are not scientifically validated.

References

[1] C. Siagian and L. Itti, "Biologically inspired mobile robot vision localization," IEEE Transactions on Robotics,vol.25,no.4,pp. 861–873, 2009.

[2] J. Li, M.D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no.4, pp.996–1010, 2012.

[3] Y Su, Q. Zhao, L. Zhao, and D Gu, "A brupt motion tracking using a visual saliency embedded particle filter," Pattern Recognition,vol.47,no.5,pp.1826–1834,2014.

[4] C.Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," in Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08), pp.1–8, June2008.

[5] X. Houand L. Zhang, "Thumbnail generation based on global saliency," in Advances in Cognitive Neurodynamics—ICCN 2007, pp.999–1003, Springer, Amsterdam, heNetherlands, 2007.

[6] A. Borjiand L. Itti, "State-of-the-art in visual attention modeling," IEEE Transactions on Pattern Analysis and Machine Intelligence,vol.35,no.1,pp.185–207,2013

[7] L. Itti and C. Koch "Computational modeling of visual attention" Nature Reviews Neuroscience, 2(3):194-203, 2001.

[8] A. Borji, D. N. Sihite, and L. Itti, ``What/where to look next? Modeling top-down visual attention in complex interactive environments," IEEE Trans. Syst., Man, Cybern., Syst., vol. 44, no. 5, pp. 523 538, May 2014.

[9] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?" in CVPR, 2004.

[10] B. Suh, H. Ling, B. B. Bederson, and D. W. Jacobs, "Automatic thumbnail cropping and its effectiveness," in Proceedings of the 16th annual ACM symposium on User interface software and technology. ACM, 2003, pp. 95–104.

[11] L.Marchesotti, C. Cifarelli, and G. Csurka, "A framework for visual saliency detection with applications to image thumbnailing," in ICCV, 2009.

[12] J.Wang, L. Quan, J. Sun, X. Tang, and H. Shum, "Picture collage," in CVPR, vol. 1. IEEE, 2006, pp. 347–354.

[13] Constantinidis, C., and Steinmetz, M. A.: Posterior parietal cortex automatically encodes the location of salient stimuli. The Journal of Neuroscience 25(1), 233-238 (2005)

[14] Ardizzone, E. and Bruno, A. and Mazzola, G., "Visual Saliency by Keypoints Distribution Analysis", Image Analysis and Processing ICIAP 2011 pages (691--699), 2011.

[15] Laptev and T. Lindeberg. Space-time interest points. In International Conference on Computer Vision, 2003.

[16] Shahzad Anwar, Qingjie Zhao, Muhammad FarhanManzoor, and Saqib Ishaq Khan, "Saliency Detection Using Sparse and Nonlinear Feature Representation" Accepted 11 March 2014.

[17] Yuming Fang,, Guanqun Ding, Jia Li, and Zhijun Fang, Deep3DSaliency: Deep Stereoscopic Video Saliency Detection Model by 3D Convolutional Networks", IEEE Transactions on Image Processing, Volume: 28, Issue: 5, May 2019, DOI: 10.1109/TIP.2018.2885229.

[18] YUZHEN NIU, JIANER CHEN, XIAO KE, JUNHAO CHEN "Stereoscopic image saliency detection optimization: A multi-cue-driven approach"2018 IEEE. Translations, DOI 10.1109/ACCESS.2019.2897404.

[19] Rajib Kumar Jha "INTERNAL NOISE-INDUCED CONTRAST ENHANCEMENT OF DARK IMAGES", 978-1-4673-2533, IEEE 2012.

[20] Rajlaxmi Chouhan "Wavelet-based Contrast Enhancement of Dark Images using Dynamic Stochastic Resonance" CVGIP '12, ACM 978-1-4503-1660, 2012.