Saliency maps detection for stereoscopic video based on the combination of visual Saliency, motion saliency, and depth Saliency

Praveen Verma[1], Abhay Bhatia[2], Parag Jain[3], [4]Bhuvenashwer Swaroop, [5]Gaurav Chaturvedi
[1]Shri Ram Group of colleges, Muzaffarnagar, Uttar Pradesh, India
[2] Dreams College of Polytechnic, Saharanpur, Uttar Pradesh, India
[3,4,5,6] Dept. of CSE, Roorkee Institute of Technology, Roorkee, Uttarakhand, India
[

Abstract

The built-up of the system is used to calculate the stereoscopic video highlight map based on a combination of visual, motion, and depth enhancements. And these three representation maps are dynamically combined into the final representation map. The working of a system can be made more efficient by reducing the three highlighting techniques to one system. The graph-based approach is used to calculate visual emphasis. GBVS is used to efficiently run global competition between targets and distractors in complex scenes. GBVS shows a higher representation in the center of the image plane and more consistently infers the human gaze than other standard algorithms. The main application of this system is to use saliency maps to identify pirated content, while other applications are image/video compression, object detection and detection, and readdressing of images. The system also needs to recognize blurred areas in stereoscopic images.**1. 44**

Introduction

In the past decades, the field of computer vision, which refers to the strategies for acquiring, processing, analyzing, and expertise the visual world, has invoked many studies interests. For the development of computer vision system as per description available the one appropriate technology is to use the duplication of abilities of the human vision system.

When talking about Saliency, the idea about the attention should be considered the topmost priority. Attention and Saliency are strongly correlated in each neurological structure and psychological definitions. The allocation of cognitive resources to information can be referred to as attention. We may explain this idea by looking at a bigger image with rich contents and determining that it is highly hard to observe information from all regions of the image at the same time. As aforementioned, we may also begin from the maximum conspicuous a part of painting after which, after a short length of time, shifts our focus to new locations (as shown in Fig 1.3).

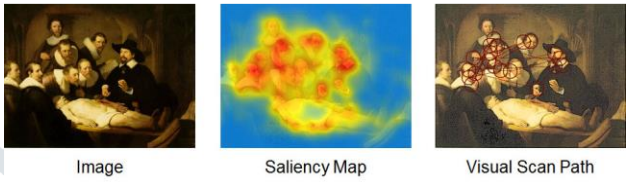


Fig. 1 Visual Saliency has an impact on human gaze policy. Images are chosen from the benchmark that has been proposed. [15].

Visual Saliency notices the salient location, feature, or salient object. The primary aim of visual Saliency is to measure the importance of various visual subsets popping out the targets and inhibiting the distractors. Several mechanisms can be observed during the computation of visual Saliency.
Here we will describe several bottom-up mechanisms:

1. **Location-based category:** In this section, Saliency is implicit to be assigned to specific locations (e.g., pixels or macro-blocks).
2. **Object-based category:** In this section, Saliency is directly assigned to objects.

Out of all, one characteristic of the bottom-up approaches is that of cues. Which are mainly from the input scene, will be a consideration when computing Visual Saliency.

A saliency map reveals the saliency subsets in an image or video frame. As shown in fig. 2, any such saliency map may be shown in two different ways. To begin, the saliency map seemed to be a grayscale map, with each pixel lying within the dynamic range of [0, 255]. A pixel with a higher gray value is considered to be more prominent. For location-based Saliency, computation is based on higher gray values form of Saliency. In the second form, a saliency map is shown as a binary mask with 1 for the salient pixel and 0 for the background pixel. This shape of saliency map is generally used for object-primarily based totally saliency computation.

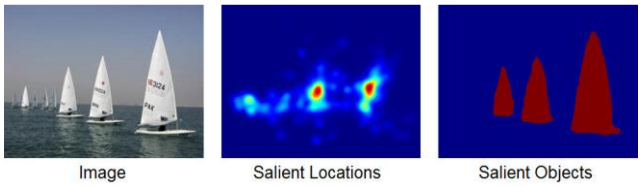


Fig. 2 Location-based and object-based visual saliency maps [2]

For the computation of visual Saliency, there are the following criteria to differentiate between bottom-up and top-down saliency models:

- **Bottom-up saliency model:** By combining pre-attentive features, we Calculate using the Bottom-up saliency model without any preconception. The feature integration is dependent on cues from predefined assignments or prior knowledge.
- **Top-down saliency model**: The "feature saliency" mapping in a top-down model is guided by cues from uncoerced assignments or obligatory priors learned during training phases.

The number of theaters that sustain 3D content has increased incredibly. Stereoscopy is different from 3D because three dimensions (Depth) are only the simulated, but Stereoscopic imaging requires at least two images, simulating our two eyes. 3D- Images can be shown as stereo pairs in 3D – Viewer or on a computer scene.

Almost every among us has seen a movie on a television screen (CRT/LCD/LED/OLED/PLASMA) or in any multiplex or cinema theatres and if they are not equipped with 3D screens, then it is required to wear 3D glasses as it gives impressions such as things are popping out of this screen or it can define the depth of the content we are seeing.

Stereoscopy: It is a technique of enhancement or creation for the illusion in the form of depth in any image with the help of using stereopsis for binocular vision. The majority of stereoscopic techniques show two offset images that are independent as seen to the left or right eye of the one who viewed the image. Stereoscopy also creates the illusion of 3-D depth from given 2-D images [1].

Stereo video, or stereoscopic video, produces the illusion of a 3D image in moving form. Various methods are used to achieve this effect; usually, a viewer needs to wear glasses. Stereo video is based on the same principles as 3D image. A picture is displayed to the spectator that combines two pictures, one for each eye.

2. Related Work

Visual saliency computations' prime objective is to describe the measuring regions, which might be specially to human observers. The majorities of image saliency detectors are mathematical or biologically inspired [1]. Several approaches have been presented in an attempt to imitate the visual system features of any human being. Most techniques assume that an object is salient if it varies significantly from other objects in its general neighborhood. [5].

To show the difference between various saliency detection techniques, first, distinguish between top-down and bottom-up approaches [10]: Top-down approaches identify important objects like faces, whereas bottom-up approaches analyze the pixel values and compute saliency values for each pixel like high pixel contrast and group highly salient pixels into regions.

A system was presented by Dittrich and Kopf [13] that can automatically detect salient image regions in stereoscopic videos. This given system is based on three major saliency detection techniques, which function with individual frame colours, information from a stereoscopic camera, as well as depth saliency and object detection. The above-defined components are dynamically combined into one final saliency map. Hence such a combination allows for a more efficient algorithm.

Itti *et al.* [12] presented a biologically inspired saliency model. They suggested using a set of feature maps composed of three complimentary channels: intensity, colour, and orientation. The overall saliency map was created by linearly combining the normalised feature maps from each channel. Even while this model has been demonstrated to accurately predict human fixations, it is relatively ad-hoc in that it has no objective function to optimise and many parameters must be adjusted by manually.

The bottom-up approach was proposed by Ma and Zhang [11], which was a contrast-based saliency detection technique for generating Saliency that operates at any single scale. A scaled and colour quantized CIELUV picture is used as the input of such a locally contrast-based map; moreover, it is sub-divided blocks that are made up of pixels. The saliency map is made by adding the differences between picture pixels and their corresponding surrounding pixels in a local area. This framework extracts the regions of attention and their points. The method which is applied to the contrast map to identify salient regions from the saliency map is termed fuzzy growing.

A new bottom-up visual saliency model proposed by Jonathan Harel et al. was based on Graph-Based Visual Saliency (GBVS) [1]. This model consists of two phases: the first one is the creation of activation maps while the help of feature channels has been taken, and then comes the normalization of those created maps. The key point of GBVS is that it shows higher Saliency at the center of the image “plane," and it guesses human fixations more consistently than other standard algorithms.

Goferman et al. [13] proposed a way in which the regions near the salient objects that may be useful in providing context are included besides finding salient areas. This method is supported by four human visual attention basic principles: local low-level considerations(contrast and color), global considerations (which suppress frequently occurring features while preserving features that deviate from the norm), visual organisation rules (which state that visual forms may have one or more centres of gravity around which the form is organised), and high-level factors (such as faces).

For detecting moving objects in images or videos, one of the most famous techniques is background subtraction. Background subtraction is used in many applications, such as traffic monitoring, video surveillance, and human motion capture. This algorithm is very fast, flexible, and precise in

terms of pixel accuracy [6]. Background subtraction is a technique used for the detection of moving regions during subtraction or differentiation of the current image from its referenced native image.

For the computation of disparity maps, the semi-global matching (SGM) technique is used and developed by Heiko Hirschmuller [9]. The method creates a decent environment balance between runtime and accuracy, especially at object borders and fine structure. It uses a pixel-wise matching cost for compensating radiometric differences of input content (i.e, image), and it performs fast approximation by optimization from all directions of incoming paths

Hao Cheng, Jian Zhang, Qiang Wu, Ping An & Zhi Liu [16] a proposal have been made of a visual saliency prediction model for any stereoscopic image that is based on stereo contrast and stereo focus models. The stereo contrast model is purely based on the color/depth contrast and moreover the pop-out effect rather if we talk about the stereo focus model, than it describes the degree of focus generally based on monocular focus and the comfort zone

Proposed System

An overview of the proposed saliency system can be understood in Figure 3. Initially, the first takes stereo video as an input and then splits this video into frames (left and right views). On the basis of such structures that are used to create a disparity map. We make the assumption; all input stereo videos are filmed with the standardized cameras and the visuals have already been corrected.

The salient regions are only detected when several saliency detection techniques are applied to the images (left and right). Each detector out of three used for salient regions creates an individual saliency map. After which, all their saliency maps are fused into a single map which has an associated weight of each map differently. The final saliency map, which contains the final salient region, is created by combining the results. Now, this information is applied to the other applications too so that enhancement of the processing of stereo videos can be done.

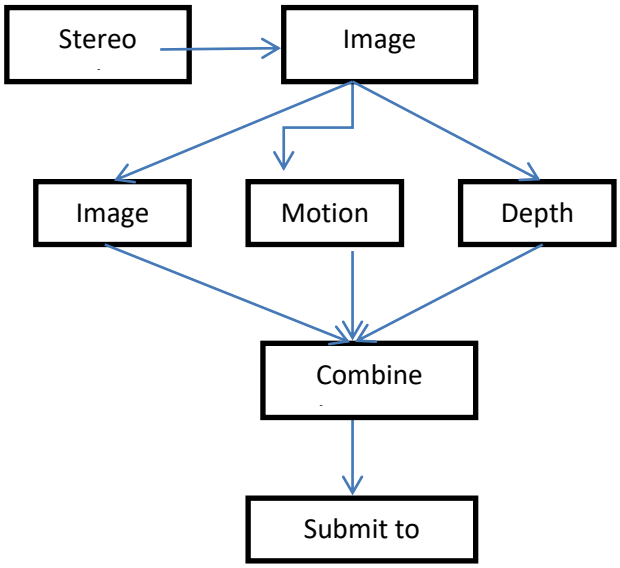


Fig 3 : Overall architecture of proposed work

System Overview

4. Saliency Detection

4.1 Saliency in still images

Graph-based algorithm (GBVS) is used to calculate the saliency of any picture suggested by J. Harel [1]. This method's fundamental notion is based on intuitive observations. As illustrated in fig. 4, an image may be represented as a fully linked network with nodes corresponding to visual subsets of various sorts (e.g., macro-blocks). The edges are weighted based on their dissimilarities and the proximity of subgroups between them. On the graph, a random walker is utilized, with the dissimilarities controlling the transition from one node to the next. Less-visited nodes might gradually emerge throughout this random walking process because they are distinctive or unusual in a global context.

As illustrated in fig. 4, an image may be represented as a fully linked network with nodes corresponding to visual subsets of various sorts (e.g., macro-blocks). The edges are weighted based on their dissimilarities and the proximity of subgroups between them.

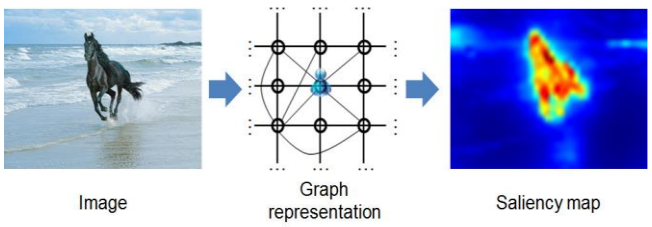


Fig.4 Representation of image as a graph for detection of less-visited salient nodes [2]

Three important phases are necessary when estimating visual saliency using a graphical model, which are as follows:

1) Extraction of features
2) Construction of various graphs
3) Saliency computation

Among the steps defined, the first one is easily accomplished by extracting the pre-attentive features (average based) in each macro-block. Intensity, hue, orientation, and even temporal aspects like motion and flicker are examples of these properties.

Assume that there are totally K pre-attentive features. The dissimilarity of visual subsets located at M (i, j) and M (p, q) may be determined as when just the kth feature is considered.

$$d\big((i,j)\square(p,q)\big)\square\left|\log\frac{M(i,j)}{M(p,q)}\right|$$

The definition for dissimilarity is undirected. For some of our experiments, the logarithmic dissimilarity can be replaced with | M(i, j) – M(p, q) |. The closeness of two content gathered (images taken from videos) should also be considered in determining the edge weight w ((i, j), (p, q)) by an exponential weighting schema:

$$w\big((i,j),(p,q)\big)\square\, d\big((i,j)\square(p,q)\big).\exp\left(-\frac{(i-p)^2+(j-q)^2}{2\sigma^2}\right)$$

Where σ is a free parameter that is used for approximation of the map width, finally, a fully-connected graph image can be represented corresponding undirected nodes for visual subsets, and edges are weighted by dissimilarities and proximity.

To generate this Markov chain, the weights of the outer edges of each node may be normalized to have a total of 1. Edge weights are transition probabilities for every state, and the approach given employs nodes that may be seen as states. If a random walker walks through the graph, the proportion of time he spends in each state may accurately reflect node saliency, and irregular nodes will be distinct from the other nodes, which will be visited less frequently.

The Markov chain is equivariant, and its equilibrium distribution exists entirely, since the nodes in any network are extremely tightly coupled and it is feasible to transit from one state to another in small steps. In this experiment, the equilibrium distribution may be established by iteratively multiplying the Markov matrix with its originally uniform vector, and after n rounds, the principal eigenvector of the matrix can be constructed.

Figure 5 shows two original left and right images are taken from the stereo video, as well as the GBVS map for each. Figure 1's third image is the left image with some blanked areas produced from the GBVS map, which covers 75% of the whole image. The figure of 75% was chosen based on heuristics as the most effective proportion for decreasing background scenery without compromising the image's prominent object. In other circumstances, the user might easily alter the proportion as required.

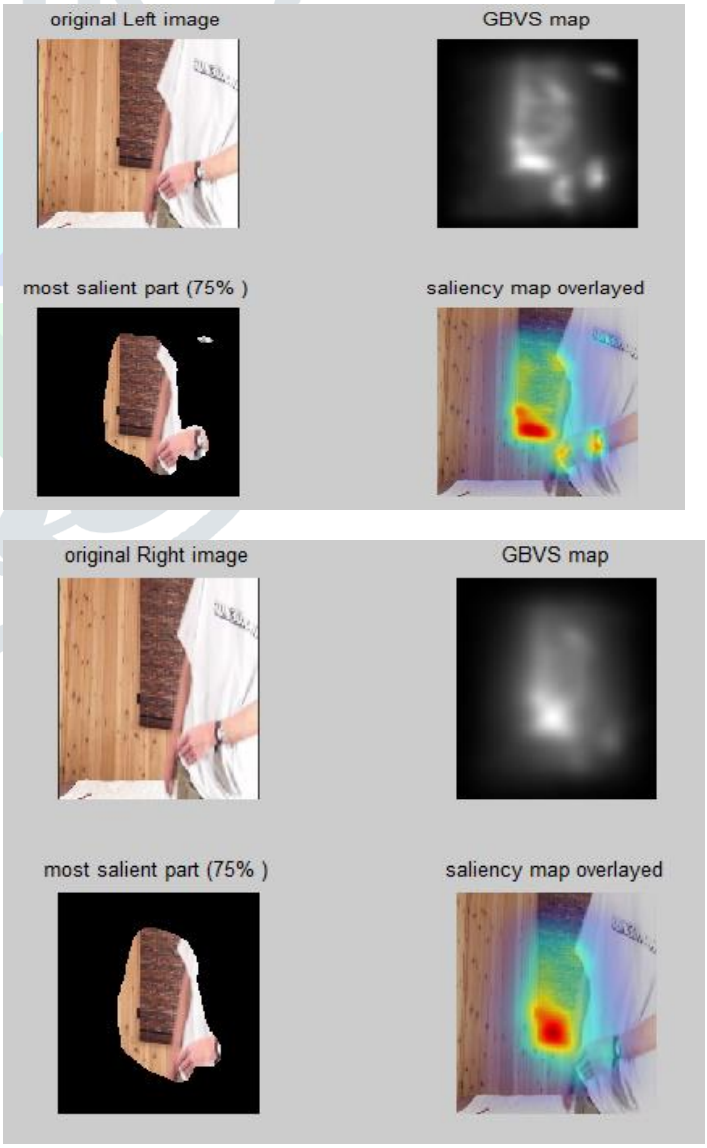


Figure 5: Two original left and right images from stereo video and corresponding GBVS maps.

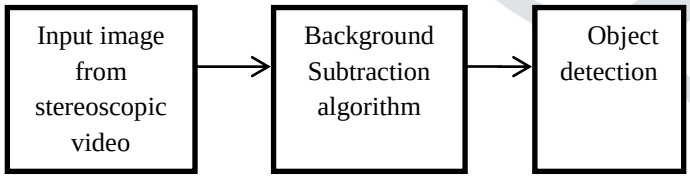
4.2 Motion Saliency

For detecting moving objects in images or videos, one of the most famous techniques is background subtraction. Many applications employ background subtraction, including traffic monitoring, video surveillance, and human motion capture. This algorithm is very fast, flexible, and precise in terms of pixel accuracy [6].

Background subtraction algorithm:

Background subtraction is used to detect moving objects from a static camera. To retrieve the moving item in the picture, this approach will remove the background from the frame. By subtracting or differencing the current image from a reference background image, background subtraction is used to detect moving areas. The background subtraction method works by first creating a background and then subtracting the current frame, which contains the moving item, from the background frame to identify the moving object. This approach is very basic and easy to grasp, and it reliably extracts target data characteristics, but it is sensitive to changes in the external environment, thus it is only suitable if the background is known. [7][17].

Two images, preferably of the same size, are extracted from stereoscopic video for motion detection. In such case, one picture is set as the background image with a moving object, while the second image is set as the current image. In addition, each image has two models. One is in the foreground, while the other is in the background. [5].



Motion Detection Model

In the foreground model moving object is present and in the background, the moving model object is not present. The first step for computing motion detection is image initialization. In image initialization, the background image is initialized. As a result, background image initialization is required for motion detection preprocessing. After performing preprocessing on each frame, a mean filter is applied to reduce the image's noise. After the preprocessing, these frames are sent into the background removal method. That subtracted image is then segmented using thresholding. In addition, Matlab was used to create the background subtraction technique.

4.3 Depth Saliency

Stereo matching is a technique for locating corresponding pixels in a pair of images, allowing for 3D reconstruction via triangulation with the cameras' intrinsic and extrinsic orientations [8]. For the computation of disparity maps, the semi-global matching (SGM) technique is used and developed by Heiko Hirschmuller [9]. The method creates a decent environment balance between accuracy and runtime, especially at object borders and fine structure. It compensates for radiometric discrepancies in input photos using a pixel-by-pixel matching cost, and it performs fast approximation by pathwise optimizations from all directions.

4.4 Fusion of the Saliency Maps

Figure 6 shows the result of the proposed map; when applied to 3D image, the most matching parts are marked and the overlay image shows where the left and right images overlay each other to give a 3D effect.

Figure 6: The original image corresponding with the Proposed Image

5. Experimental Results

Figure 7 displays the saliency maps for still images, motion, and depth information in a single frame. The stereoscopic video's image saliency map, displayed on the left side in figure 6(b), discovered the majority of the salient regions.


(a)

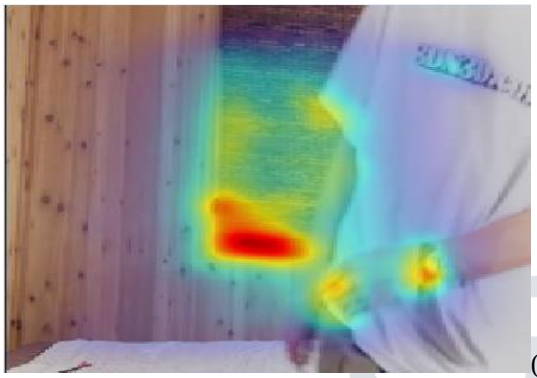

(b)

Figure 7: Left frame of stereo video (a), Image saliency (b)

6 Conclusions

In this paper, the saliency detection technique that was used in the early era of this technology has been reviewed as it can be used for proper 2D images, and for any particular object, motion detection can be done after having a cam-captured image. Rather than extending the same, we have introduced the 3rd dimension for implementation and detection of motion in 3D images too. According to the overall observation, the proposed method mostly has comparatively better results on stereoscopic images as compared to other standard methods. The characteristic that is common to these images is generally having a small number of salient areas, which can be caused by either 2D salient features, depth features, or motion features. Therefore, the saliency map generated is based on either 2D salient features, depth, or motion. These might predict parts of the salient area, but not all of them. This can be the reason why 2D maps on Saliency, depth saliency, and motion Saliency have comparable performances, but if we compare their combination than will find that the results are much better than the others.

7. References:

[1] Harel, J., Koch, C., Perona, P.: Graph-based visual Saliency. In: Advances in NeuralInformation ProcessingSystems (NIPS), pp. 545–552 (2007)

[2] Jia Li Wen Gao“Visual saliency computation” book.

[3]S. He, J. Han, X. Hu, M. Xu, L. Guo, and T. Liu. Abiologically inspired computational model for image saliency detection. In Proceedings of the 19th ACMinternational conference on Multimedia, MM’ 11,pages 1465–1468, New York, NY, USA, 2011. ACM.

[4].M. Kalpana Chowdary , S. Suparshya Babu , S. Susrutha Babu , Dr. Habibulla Khan‖FPGA Implementation of Moving Object Detection in Frames by Using Background Subtraction Algorithm‖ International conference on Comm.unication and Signal Processing, April 3-5, 2013.

[5]. Mahamuni P. D, R. P. Patil, H.S. Thakar. Motion Object Detection using Background Subtraction Algorithm using Simulink, IJRET (International Journal of Research in Engineering and Technology) 6 June 2014.

[6] M. VAN DROOGENBROECK, and O. PAQUOT. Background Subtraction : Experiments and Improvements for ViBe. In Change Detection Workshop (CDW), Providence, Rhode Island, 6 pages, June 2012.

[7]. Motion and Feature Based Person Tracing in Surveillance Videos‖ transactions on computer vision 2011.

[8] Heiko Hirschmuller, Oberpfaffenhofen, “Semi-Global Matching – Motivation, Developments and Applications”

[9] Hirschmüller, H., 2005: "Accurate and Efficient StereoProcessing by Semi-Global Matching and Mutual Information",IEEE Conf. on Computer Vision and Pattern RecognitionCVPR'05, Vol. 2, San Diego, CA, USA, pp. 807-814.

[10] V. Navalpakkam and L. Itti. An integrated model oftop-down and bottom-up attention for optimizing detection speed. In Computer Vision and PatternRecognition, 2006 IEEE Computer Society Conferenceon, volume 2, pages 2049 – 2056, 2006.

[11] Y.-F. Ma and H.-J. Zhang. Contrast-based imageattention analysis by using fuzzy growing. In Proceedings of the 11th ACM international conferenceon Multimedia, pages 374–381. ACM Press, 2003.

[12] Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapidscene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence20(11), 1254–1259 (1998),

[13]S. Goferman, L. Zelnik-Manor, and A. Tal. “Context-aware saliency detection”. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEEConference on, pages 2376 –2383, june 2010.

[14] T.Dittrich, S. Kopf, W. Effelsberg: “ Saliency Detection for Stereoscopic video”. 4th ACM Multimedia system conference (MMSYS), pp. 12-23, february 2013

[15] Li, J., & Gao, W. (Eds.). (2014). Visual saliency computation: A machine learning perspective (Vol. 8408). Springer.

[16]. **Hao Cheng, Jian Zhang, Qiang Wu, Ping An & Zhi Liu Stereoscopic visual saliency prediction based on stereo contrast and stereo focus EURASIP Journal on Image and Video Processing volume 2017, Article number: 61 (2017)**

[17] Jain, P., Vijay, S., & Gupta, S. C. Performance Analysis & QoS Guarantee in ATM Networks. *Global Journal of computer Science and Technology*, 131-136.