https://doi.org/10.33472/AFJBS.6.Si3.2024.528-542

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

Detection of Images for Video based Stereoscopy by Using the Combination of Saliency Methods

Beerbal Solanki[1], Updesh Kumar Jaiswal[2], Bhupesh Kumar Gupta[3], Manoj Kumar[4]

[1,2,3,4]Assistant Professor, Department of Computer Science & Engineering, Ajay Kumar Garg Engineering College, Ghaziabad

Email: [1]solankibeerbal@akgec.ac.in,[2]jaiswalupdesh@akgec.ac.in, [3]guptaindia81@yahoo.co.in, [4]mk.miet@gmail.com

Article Info

Volume 6, Issue Si3, May 2024

Received: 08 April 2024

Accepted: 03 May 2024

Published: 30 May 2024

doi: 10.33472/AFJBS.6.Si3.2024.528-542

ABSTRACT:

Uses of video-based spectroscopy are expanding in many domains where depth information is crucial, including 3D vision, medical diagnostics, distant object identification, and space reconnaissance. For accurate image processing and to enhance the authenticity, depth information is vital in all of these cases. Saliency detection is a useful technique in this method for predicting which parts of images or movies people would find most interesting. However, stereoscopic visual saliency has received surprisingly little attention. This study investigates and evaluates the most recent multimodal fusion based picture identification model from stereoscopic video coupled with saliency approaches in an effort to address this deficiency. The results demonstrate that the adaptive multimodal fusion based spectroscopic enabled image identification model with saliency features is superior when compared to an image detection model from stereoscopic movies that does not use saliency approach.

Keywords:3D video spectroscopy, image detection spectroscopy, saliency method, adaptive fusion spectroscopy, non-saliency

1. Introduction

Among the most basic and difficult jobs in image and video analysis is object detection. A lot of people have been thinking about how to improve computer vision for tasks like object categorization, object counting, and object monitoring recently. Aside from the grey value information that describes intensities, colours, densities, etc., numerous image processing applications also need explicit depth information.

There are several examples of such applications in different industries, including 3-D vision (remote sensing systems, robot vision, photogrammetry), medical imaging (computer tomography, microsurgery, magnetic resonance imaging), and visual communications aiming to achieve virtual presence (virtual travel, conferencing, education and shopping, virtual reality).

Depth information is crucial in all of these instances, since it is necessary for precise picture processing and for augmenting the authenticity(Lagendijk et al., 2009).In this approach, saliency detection is the process of predicting the specific spots in pictures or videos that capture the attention of humans. Several static picture saliency detection methods have been developed in recent years. Nevertheless, there has been relatively limited exploration of stereoscopic visual saliency.

A noteworthyaspectto give rise to this outcome is the absence of a comprehensive benchmark dataset that includes an adequate number of video sequences with accurately matching depth information. The information derived from the depth map is essential for accurately detecting saliency. Firstly, stereoscopic data, as opposed to two-dimensional (2D) data, is more suited for practical implementation. Furthermore, with the increasing complexity of visual scenes, relying just on 2D data is insufficient for extracting prominent items. The acquisition of a depth map enhances the ability to distinguish between several things that have a similar look in intricate settings.

As an attempt to resolve this shortfall, this article explores and analyses the evolving video based stereoscopy approaches integrated with saliency methods for image detection. In particular, this study pays special attention to adaptive fusion based 3D image detection from saliency included video based stereoscopic techniques. To justify the efficacy of the focused model, a stereoscopic approach without saliency method is compared with the said adaptive image detection model of stereoscopic model.

2. Related Works

Here we give a detailed overview of the existing academic literature on the topic of spectroscopy-based image recognition with saliency approaches, including its function, purpose, and importance. Several methods are explored in order to get understanding of the best model for video spectroscopy-based picture identification, with saliency characteristics being crucial in obtaining depth data. Using better tactics that are all described in this paper, AI models are presently being deployed in this domain.

While evolving in its application area of image detection, saliency detection algorithms, as argued by (Fang et al., 2013) have been extensively utilized in many 2D multimedia processing applications. In their study, the authors studied that the growing uses of stereoscopic display necessitate novel saliency detection algorithms for stereoscopic pictures. The detection of saliency in stereoscopic images requires the consideration of depth characteristics, which is not the case for 2D pictures.

A fresh approach for the detection of stereoscopic saliency is introduced in this study. The technique extracts four distinct kinds of aspects—color, luminance, dimension and texture—

from DC-T coefficients, and this describes the energy for image patches—through the analysis of variation in the impact hue, texture, and depth aspects. A novel fusion technique has been developed to merge the feature maps in order to calculate the ultimate saliency map for stereoscopic pictures.

Later on, these observations are further enhanced. Using insights into the human visual system (HVS), it is shown in many researches, such as, the research of Fang et al. (2016) that suggests a novel saliency-based stereoscopic picture retargeting approach. Low-level characteristics like intensity, color, texture, and depth were used to develop the saliency detection approach. Stereoscopic visual attention modelling also uses the center and depth biases already present in the HVS. When scaling a picture, the saliency and edge maps are merged to determine which pixels are most visually important. When mounting a photo, the structure and depth distortion may be minimized by constructing a dynamic energy function.

Predicting where in an image or video a human viewer's gaze will be drawn is the goal of visual attention models (VAMs). More efforts have been needed to develop 3-D VAMs, and large-scale video saliency prediction datasets still need to be included. This research presents a huge volume collection of data on eye-movements culled from stereoscopic 3-D films of about 61 in number, using data from 24 people who viewed the videos freely. To validate the efficacy of current 2-D and 3-D variations of VAMs, we conduct an online benchmark evaluation of their performance on the suggested dataset. The results showcasing the currently evolved 2-D and 3-D forms of VAMs are shown on a website, and new 3-D VAMs are easily added.(Banitalebi-Dehkordi et al., 2016)

The study examines the advantages of incorporating 3D Visual Attention Models (VAMs) in Full-Reference (FR) along with No-Reference (NR) output attribute evaluation measures used to standardize stereoscopic 3D video. The research employs a sizable library of stereoscopic movies with varying distortions to assess the effectiveness of saliency maps in improving the quality evaluation task for stereoscopic video. To achieve this goal, authentic 3D VAM platforms are combined with several previous FR and NR stereoscopic video standard measurements in a quality evaluation pipeline. According to the results of the tests, applying stereo saliency, in general, increases the accuracy of the quality rating, and this increase is much more pronounced when assessing the quality of NR videos. As reported (Banitalebi-Dehkordi, Nasiopoulos, 2018),

As inferred in various experimental findings that most retargeting frameworks use graphical attention models to identify the scene's salient area, although few such models are available in the literature. Prior studies have demonstrated that prominent positions in three-dimensional objects may originate from depth characteristics that are disregarded in illustrations of two-dimensional graphical attention. Consequently, it is difficult to create and incorporate novel visual attention models for three-dimensional objects into the stereoscopic image retargeting framework.

However, some models treat depth perception as a distinct visual channel, while others rely on the more simplistic notion of viewing bias. Utilizing a stereoscopic image eye-tracking dataset, the hybrid model put forward in this study is validated. This model takes into account various stereo seeing biases and 2D/depth attributes that have been documented in previous psycho-visual research. The results of the experiments show that the proposed approach works better than the ones that are already out there in this field. (Wang et al., 2014).

It has been shown experimentally that stereo saliency detection and picture retargeting can outperform similar technologies now available in the public domain. We could quickly expand the suggested stereoscopic image retargeting approach by including the top-down saliency maps of the individual objects in our saliency detection model(Chen et al., 2015).

Because the HVS is so attuned to visual distortions on the human form, we may use a saliency model to determine which pixels in a resized picture are most important.

With the advancements in image capture technology, saliency detection techniques for 3D pictures have been created. Most 3D saliency detection strategies use deep contrast or global depth priors to boost performance. The authors offer a method for directly extracting local background characteristics using depth data and multi-layer RGBD models. Using a foreground super-pixel dictionary, Li et al. got more precise foreground saliency data.

At first, Wang et al. used lines and 115 stereoscopic variations as discriminatory saliency cues to set a saliency leaning for the picture pairs. Right after that, Zhu et al. proposed an FCN method that has a main system that works with RGB values, a secondary network that uses depth impulses a lot, and adding depth-based traits to the main the system. In order to discover the interaction mechanism using low-depth and low-color features, Qu et al. developed a CNN. Riku et al. developed a CNN framework that is deep by incorporating depth properties. (P. Zhang et al., 2020)

A video saliency approach was put together by Wang et al. that utilizes convolutional networks. This model integrates the spatial saliency estimate into the dynamical saliency approach, enabling the immediate generation of spatiotemporal saliency inference. This eliminates the time-consuming optical flow computation. Stereoscopic display technologies have increased the significance of binocular research on visual saliency(Cheng et al., 2017).

Bruce et al. introduced a stereo saliency model that generalizes the standard 2D attention model to the binocular realm. For adaptive rendering, a method for extracting an area of interest (ROI) was also described. Using disparity information to weight a 2D saliency map, Chamaret et al. (2017) generated a saliency map of a stereoscopic image.

Potapova et al. created a stereoscopic saliency recognition approach by adding information from the top down to the usual bottom up method. Eye-tracking software was used with both 2D and 3D photos to figure out how salient a depth was. They were able to make a binocular saliency map because of this. For instance, how stereo salient a picture area was was based on how far away it was from a comfortable zone.

Calculating contrast between colour, intensity, texture, and depth features was recommended by Fang et al. to generate saliency maps produced by stereoscopic methods. Zhang et al. gave an approach of a visual based saliency model for 3D pictures developed on deep learning-based platform. Pretrained convolutional neural network models were used to extract colour and depth information to make saliency predictions.

After using Kim et al.'s recommendations, a saliency map is constructed utilizing low-level features, motion and depth characteristics, and high-level scene type. Reference: (Zhang et al., 2019). Saliency prediction models for stereoscopic films are proposed in this work, including one that draws inspiration from component-based interactions to learn to explore saliency. To represent the saliency caused by spatiotemporal coherence between successive frames, the model initially uses a 3D residual network (3D-ResNet).

When you use a deep convolutional network (ConvNet), included depth naturally figures out what's important depending on how the left and right views move together. As a final step, we aggregate saliency distributions derived from many components into one final saliency map over time using a component-wise refinement network. The suggested model has shown superior performance in experiments compared to the best existing saliency-detecting methods. (Q. Zhang et al., 2020)

Numerous models have been built using deep learning to predict saliency for pictures or videos. An early visual saliency detection model was created by Vig et al., followed by The fully convolutional neural network (FCN) inspired visual focus approach by Kruthiventi et al. Huang et al. sought to make saliency forecast and human eye focus more in sync with each

other. Cornia et al.'s sight saliency paradigm was used to predict where people would look in nature photos. Wang et al.'s FCN-based video saliency model, Lai By fusing deep features in colour and depth channels retrieved from a pre-trained AlexNet model.
Zhang et al. suggested a learning-based saliency detection model for stereoscopic pictures. To identify visual attention, Nguyen et al. used seven low-level characteristics derived from stereoscopic picture pairings. Huang et al. created a learning-based model for stereoscopic films that blends 2D features and depth information into one. They used the eye fixation map as a benchmark. Yang et al. showed a two-stage grouping method to fix the problems caused by low-quality depth footage.
The research by Wang et al. (2018) uses the disparity map and discrepancy image to show a new way to find the most important parts of stereo images for a stereoscopic image quality assessment (SIQA). Using the picture data, the inter-view disparity, and the difference map, a brand-new quaternion representation (QR) of each stereo image is built. The visual saliency maps for stereoscopic picture pairs' left and right views are generated by applying the quaternion Fourier transform (QFT) on the produced QR.
Compared to existing visual saliency models suggested for stereoscopic pictures, experimental findings show that the proposed model can greatly improve SIQA's performance. This result further verifies the efficacy of the suggested visual saliency model in representing the acuity attribute of the HVS in evaluating the perceptual quality of stereoscopic images. When applied to stereoscopic pictures, the proposed QRSIVS-based SIQA measure provides reliable predictions of their 3D visual quality.
Researchers have not yet settled on an ideal approach for determining if depth data is accurate. Problems with computation and prediction arise due to the data-driven nature of deep learning techniques and the absence of a trustworthy dataset. In addition, deciding on saliency features and how to threshold them can be a challenge when dealing with complicated picture recognition and feature extraction processes. It is common to find that single model frameworks are failing; as a result, these frameworks should be improved in terms of their depth data estimation efficacies. Likewise, other options should be considered.

Research Scope and Motivation

Since 3D TV output stopped, methods for turning 2D into 3D have become more important. Because virtual reality devices that use split perception are so common and well-liked, this is the case. A lot of film footage is mostly in a two-dimensional version. In order to fully utilize the capabilities of stereo vision systems, it is crucial to develop solutions that enable quick, dependable, and flexible 3D transformations(Hachaj, 2023).
Visual attention models have several applications in image processing, including categorization, visual retargeting, visual coding, watermarking, picture segmentation, and image retrieval. The progress made in 3D display technology and gadgets has led to the creation of many new uses for 3D multimedia, such as updating 3D videos and deciding on 3D video requirements, among others. In general, the need for visual attention-focused 3D application development has led to a greater need for computer models that can figure out what information in 3D multimedia is important.The field of computer vision has generated significant interest in the identification of salient objects.
As stereoscopic displays become more popular, they may give users a feeling of perspective. This has made it more important to create computed saliency recognition approaches for 3D multimedia apps. When finding the most important parts of 3D images, unlike 2D images, depth is very important (Chacko et al., 2016).
The newly released sensing technologies, such as Microsoft Kinect, offer exceptional capability and adaptability to record RGB-D pictures. Depth, together with RGB information,

has been demonstrated to be a useful cue for extracting saliency. Ju et al. also came up with a new way to use the anisotropic center-surround disparity to find important features in depth views. For RGB-D images, saliency recognition must take into account the depth factor, which is not the case for 2D images.

Depth cues offer further crucial information on the objects present in the visual field, making them significant characteristics for detecting saliency. RGB-D co-saliency detection is a novel and intriguing topic in saliency detection. It focuses on identifying the shared salient objects in a collection of RGB-D photos, utilizing the added depth information. Stereoscopic material provides significant supplementary binocular signals that enhance human depth perception. Hence, the primary obstacles in the development of 3D saliency models are accurately gauging saliency based on depth cues and effectively integrating saliency derived from depth characteristics with saliency derived from other 2D low-level data(S. Wang et al., 2018).

These approaches are being advanced continually. Currently, the fusion approach and acquisition of adequate cross-modality complementary information are essential factors in RGB-D salient object detection (SOD). The fusing method employs a selective fusion technique to mitigate contamination resulting from incorrect depth information and efficiently integrate multi-modal information by combining unimodel features. Hence, it is important to mitigate the adverse effects of low-quality depth photographs and carefully choose dependable and precise information throughout the fusion process(Huang et al., 2021).

Early-fusion techniques specifically used both RGB and depth data as inputs and processed them together in a unified manner. Nevertheless, this fusion technique fails to take into account the disparity in distribution and distinct feature characteristics present in separate single modalities. It is challenging for a single model to accommodate multiple modalities(Sahu & Vechtomova, 2021). In contrast, the late-fusion technique involves processing the data from single modalities separately in order to generate the matching saliency maps.

After that, the focus process is used to make both maps. The main problem with this system, though, is that there isn't enough internal scrutiny across the two forms. The strong cross-modal signals are also squished and lost in the two separate stems.Before, we talked about two fusing methods that cause the mechanism of learning to stay in a specific optimal region, where it tends to focus on RGB data.

This occurs because the combination of channels leads to a decline in the effectiveness of learning, resulting in the final prediction being mostly influenced by the RGB features and not taking into account the valuable contributions of the cross-modality informative feature. In order to improve the fusion process of the depth maps, various studies have suggested middle-fusion algorithms that utilize two-stream CNNs to generate intermediate independent features.

In order to tackle the aforementioned problems, it is imperative to attend to the fusion process of cross-modality complementing and develop new adaptive multi-level deformable fusion modelsto include flexibilities.

3. Materials and Methods

This article presents the results of a research that was built using analytical and deductive methods. Improved depth estimation and corresponding models are the focus of this work, which draws on open-source materials already collected from studies focusing on saliency-added 3D video spectroscopy-based picture identification.

Using stereoscopic films combined with saliency criteria, this work aims to support and suggest a future method for improving image recognition. In order to achieve this goal, the research delves deep into multimodal fusion methodologies, demonstrating their effectiveness and highlighting their adaptive characteristics. In order to demonstrate the effectiveness of the fusion-based adaptive technique, we compare and analyse a CNN-based image recognition model that does not use saliency.

4. Findings and Discussion

In this section, two separate approaches of depth estimates are provided: (a) Using feature extraction modelto transform 2D to 3D images without integrating saliency methods, and (b) Image detection framework of saliency based Hierarchical Multimodal Adaptive Fusion (HMAF) Network.

(a) Platform setup and Experimental Outcome of image detection with feature extraction model to transform 2D to 3D images without integrating saliency methods

Presently, the prevailing method for generating depth images involves employing an appropriate deep neural network. In order to evaluate its efficacy, a Deep Neural Network (DNN) architecture (DD) is used for the purpose of depth prediction. The model is utilized in the depth map based spectroscopy based image detection procedure without adding saliency by Hachaj (2023). Again the model utilizes single separate model of feature extraction and do not employ fusion as a part of corrections and improvements of information. In this research, the researcher stressed that depth feature extractors are novel technologies that have found acceptance in many applications.

In the same way, the DIBR approach is used to predict depth pictures, that are are also called depth maps and are based on the left image. A two-dimensional snapshot, which is used to make a stereoscopic picture pair, is thought to correctly show the photo from the left camera (left side snapshot). Our objective is to accurately predict the corresponding right camera image (right side snapshot).

This study looks at how well a dense depth (DD) arrangement of filters and reconstructs depth pictures at different sizes utilizing a U-Net design and a DenseNet framework that has already been receive training. The DenseNet-169 backbone is used by DD. It is a simple and clear framework.

This backbone lacks residual connections and is based on forward connections. Consequently, the decoder may be resized by establishing skip connections with the desired resolution. An important benefit of DD is that it allows for the creation of deep estimation networks with varying numbers of weights and processing speeds, thanks to the DenseNet backbone design. Conversely, the drawback is the absence of residuals in the core structure, which to some extent restricts the network's capacity for expression. As evaluated in this experiment, the performance of DD is slower.

The model is worked based on the following mathematical theory:

The depth image and left image may be utilized to produce the right picture of the stereo pair through the application of the depth image-based rendering (DIBR) approach. The magnitude of the sensor displacement, h, is determined by the following equations in this methodology:

$$h = -t_x \frac{f}{Z} \tag{1}$$

and

$$t_x = \begin{cases} -t_c & : left - eye \quad view \\ t_c & : right - eye \quad view \end{cases} \tag{2}$$

In this picture, tc stands for the interaxial distance, f for the focal length of the picture, and Z for the convergence distance. Z represents the depth value of that particular pixel.In the right and left sides of photos, the change is made to every pixel. There are clear "holes" in places where there is a lot of variation because it is per-pixel.

For filling in these gaps, numbers from nearby places where the color has been established must be interpolated. When the method mentioned in (1) is used, it makes consequently the right and left stereo vision pair images less clear. This is the reason why certain methods, such as the one mentioned in reference, only modify one of the photos while keeping the other one unaltered. In order to implement the methodology employed in (1), it is necessary to define the parameters tc and f, which may differ based on the specific video stream being addressed. The choice of stereo vision technology may also be influenced by the viewer's desire. Considering this, we may streamline the aforementioned equations in the following manner:

$$h = -D \cdot MaxDisp \tag{3}$$

With D standing for a depth picture and numbers running from 0 to 1, and MaxDisp for the biggest difference within the left and right pictures. By using the above approach, you can change the amount of depth in the DIBR picture by changing just one setting.

When the DIBR procedure is used, pixels move from the right image to the left at the edges of regions with a large and small differences, making the visual on the right side look twisted. This distortion results in errors or gaps in the created image. As the left camera fails to detect the color information of the pixels in these regions, it is necessary to fill in these gaps. One way to do such is to use techniques from the practice of digital inpainting, which make it easier to fix small, broken parts of an image. A image and a binary map of areas are used as inputs by these algorithms to tell them which pixels need to be calculated.

The fast marching methodology and the Navier-Stokes-based system are two common ways to paint digitally. The methods listed above are gradient-supported methods that are famous for being very good at filling in large areas of empty space that are both wide and tall. However, in practical implementation, the holes created by DIBR typically manifest as vertical lines with rather small widths. Therefore, in the majority of instances, it is unnecessary to locate specific regions with a limited range of frequencies that need to be filled by the approximated pixels.

The result can be better if the process is done much more quickly. With the above approach, the mean amounts of the pixels around the slots are added to the pixels throughout the gaps to get an estimate of those pixels. The technique goes through all the empty spaces in the mask picture, which are shown by non-zero pixels, one by one until all the spaces are filled. Even with a small average pane size, all the holes can be filled with just one run of the while cycle.

Given below the algorithm used for DIBR and Inpaint operations as used in this model:

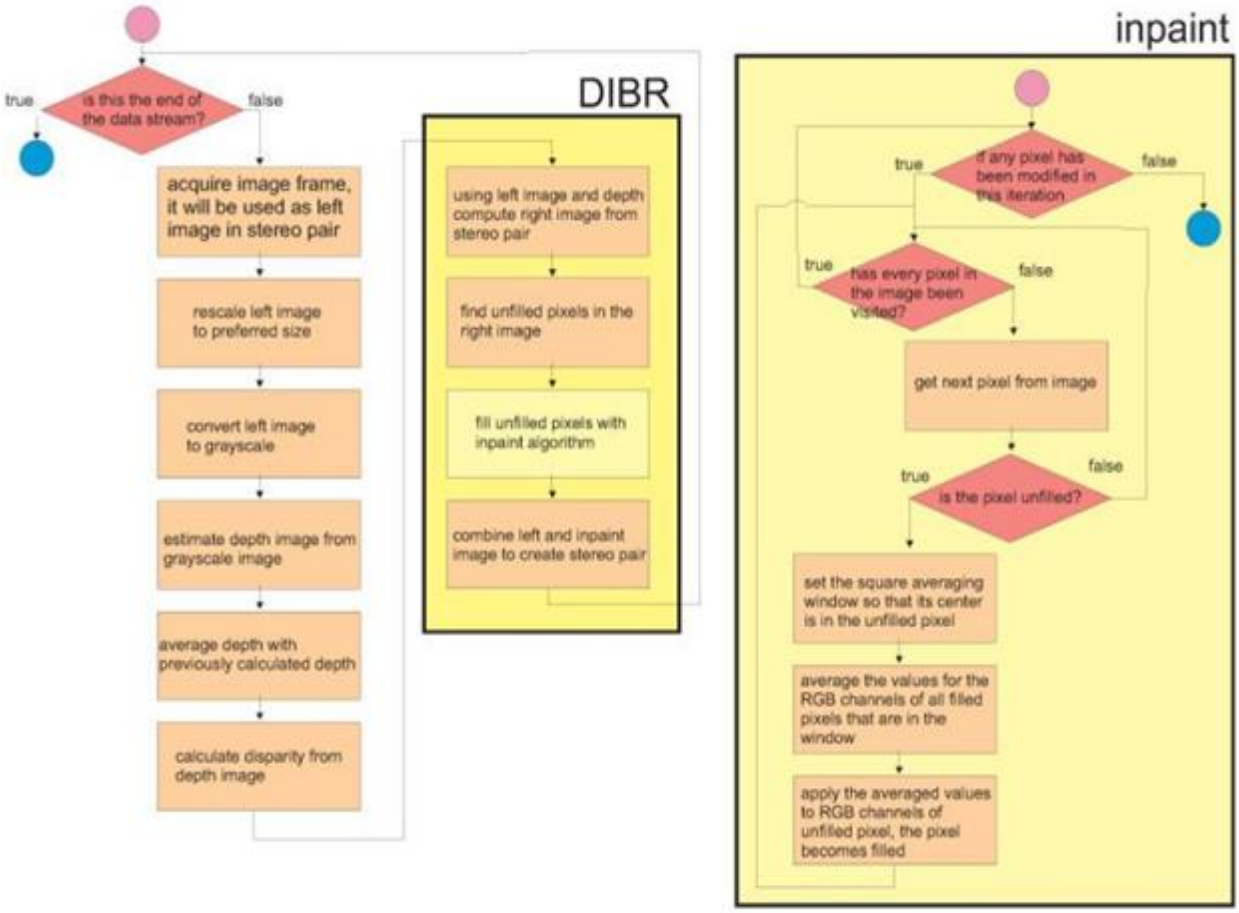

Figure: The framework flowline of DIBR and InPaint Implementation in Depth Map Based 3D Image Detection with Spectroscopy (Source: (Hachaj, 2023))

From a speed perspective, DIBR-type algorithms as used in this model encountered two bottlenecks: depth estimation and inpainting.Augmenting the maximum disparity further enhanced the three-dimensional perception of the pictures. However, the greater difference led to a reduction in the level of comfort while seeing, but this loss was not directly proportional to the increased three-dimensionality. But, by utilizing the equation presented in this study (3), users have the ability to easily adjust the maximum depth according to their preference.

However, the fact that non-DIBR-based strategies don't respond to contrast is a big problem because it means that the received picture can't be changed to fit different video recordings, stereo vision technologies, or customer options. The algorithm of Deep3D is not practical to utilize since its low-perceived rating negates any advantage in terms of speed.

(b) Platform setup and Experimental Outcome of image detection with framework of saliency based Hierarchical Multimodal Adaptive Fusion (HMAF) Scheme

In this Hierarchical Multi-scale Attention Fusion (HMAF) algorithm for predicting saliency in RGB-D images proposed by Lv & Zhou (2020), a two-stream network to extract the hierarchical characteristics of two modalities is employed. After that, three attention components (not just one feature selection model) with two input variables are used to join the two modes' hierarchical properties on the fly. In the end, the three-input attention module flexibly combines the hierarchical combining saliency features.

The working framework of the model is given below:

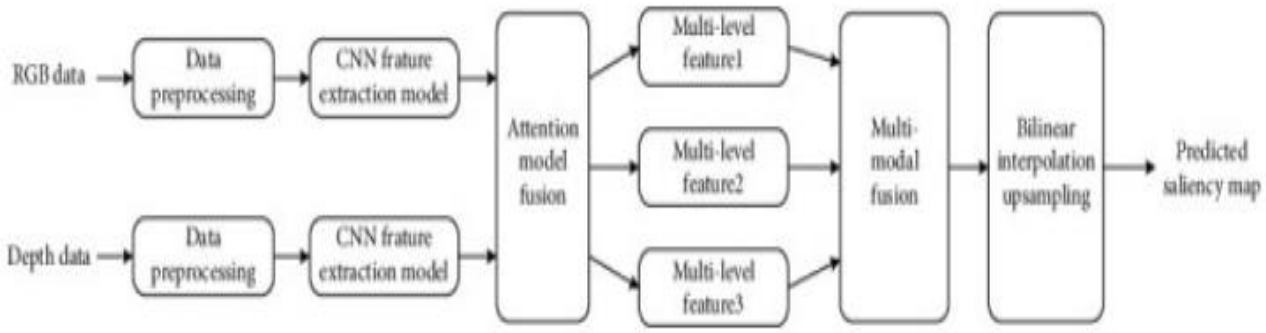

Figure: The framework of Hierarchical Multi-scale Attention Fusion (HMAF) of image detection with spectroscopy added with saliency methods (Source: (Lv & Zhou, 2020))

Features that exist at several levels have distinct characteristics. High-level characteristics provide a greater amount of semantic information, which is valuable for differentiating important locations, but they provide less spatial and local-context information. Conversely, low-level characteristics encompass a greater amount of spatial information, such as textures, edges, and curves.

Therefore, it is crucial to consider both high- and low-level variables as significant and mutually beneficial in predicting visual attention. This highlights the need for multilevel feature extraction in order to accomplish accurate saliency prediction.The study adopted VGG-16, a frequently used pre-trained network configuration, as the basis of the two-stream network for extracting ordered characteristics from the RGB and depth paradigms. The VGG-16 network was modified by eliminating its fully connected layer. In order to maintain the relatively large spatial dimensions in the upper layers, the stride of the final max-pooling layer was reduced, resulting in improved accuracy in predicting saliency.

Consequently, the network's speed was decreased to 16 for a given source RGB (depth) picture scale of $M \times N$ pixels. The resulting feature map has a dimension that is spatially equal to M broken by 24, compounded by N divided by 24. The research investigation employed the hierarchical features extracted from the layers with the greatest pooling Pool3, Pool4, and Pool5 to create organized fusion saliency feature maps for responsive multimodal a fusion procedure High-level characteristics can effectively differentiate between several item categories, but they may not provide precise discrimination among objects within the same category.

On the other hand, low-level characteristics contain greater spatial details that may be differentiated across items of the same category. But they're not as tough when it comes to drastic changes to their look. To maximize the accuracy of saliency prediction, this study included both high-level and low-level features.We retrieved hierarchical multifunctional traits from Pool3, Pool4, and Pool 5 using the HMAF-based method that was proposed.

The study has utilized two-input attention modules that consist of three operational parts: Transformation, Fuse, and Select.

Transformation module: Assuming thatfm, i, j ∈ $\mathbb{R}H \times W \times C$ relate to the feature traits of m modalities, where m can take the values 1 and 2 representing the RGB and depth modalities in terms of their precedence. The variable i can take the values 1, 2, and 3 to represent the extraction of different level features from max-pooling layers Pool3, Pool4, and Pool5 of VGG-16 respectively. The variable j represents the spatial position.

Fuse module: The results acquired from the two modal streams were fused by an element-wise summing, referred to as Fuse Transformation.

$$U = U^1 + U^2 \quad (1)$$

The channel-wise characteristics, denoted by the parameter S, which is a member of the set of real numbers that comprise ℘C, were generated by extracting and integrating the entirety of the data using the average global pooling. In addition, a compact feature Z, which belongs to the set of real numbers ℝd×1, was developed to assist in accurate and adaptable selections. The task was achieved by employing a fully connected layer, which reduced the dimensionality to improve efficiency. In this layer, the output Z is obtained by applying the rectified linear unit function (δ) to the product of the batch normalization (β) and the input matrix (WS), where W is a real-valued matrix of size d×1.

Select Module: Soft computing techniques were employed to provide a soft attention mechanism that selects alternative streams based on the characteristic Z. A two-stream probability distribution according to statistics was generated by applying a sigmoid process operation to the integers in a layer that is completely linked.

$$w_{1,c} = \frac{e^{A_c Z}}{e^{A_c Z} + e^{B_c Z}},$$
$$w_{2,c} = \frac{e^{B_c Z}}{e^{A_c Z} + e^{B_c Z}}. \quad (2)$$

The cth line of matrix A (B) is represented by Ac (Bc) in this scenario, while the cth member of vector a (b) is represented by w1,c (w2,c). In addition, the two soft attention vectors for U1 and U2 are denoted by A and B, consequently, while a and b stand for their related versions.After reassigning weights w1,c and w2,c to aspects that translate to the RGB and depth modalities, respectively, the final outputting feature map Y was produced. The equation that corresponds to this may be expressed as

$$Y = w_{1,c} \cdot U^1 + w_{2,c} \cdot U^2 \quad (3)$$

It is important to mention that the equation shown above is only valid for the scenario with two inputs. However, one may readily get the relevant formulas for scenarios requiring additional inputs by straightforwardly expanding equations.

Hierarchical fusion saliency properties were obtained after completing a procedure of fusing different feature modes. The resulting characteristics were then extended to the same criteria using bilinear interpolation.Then, via extending problems, a three-input attention module was used to obtain the eventual fusion conclusion. Earlier in Section 2.2, it was mentioned that the attention module may dynamically assign significant weights to fused saliency properties from different layers.

Using two datasets, NUS3D along with NCTU, this research project evaluates the performance of saliency-prediction methods. Here is how the analysis is carried out: (1) There are a number of 2D and RGB-D view instances included in the 600 RGB-D photos that make up the NUS3D collection. 2D and RGB-D fixation map data, as well as depth pictures and RGB triggers, are available. The NCTU dataset has 475 RGB-D pictures and depth

information. A wide variety of sequences, culled from many sources including pre-existing stereo clips and films, make up this collection.
The suggested strategy consistently outperforms contemporary single module based approaches. The combined loss functions as used as evaluation criteria yield superior information in favor of the suggested strategy when they matched to individual functions. What sets the overall loss function apart from its competitors is its ability to produce competitive forecasting outcomes within all criteria.
In order to demonstrate the influence of hierarchical properties, visual saliency is predicted using the resultant attributes from convolution layers Pool3 together with Pool4, and Pool 5.From the analysis, it may be concluded that hierarchical characteristics are significant and mutually beneficial, resulting in accurate saliency prediction.
When compared to other methodologies, the suggested Hybrid Multi-Attention Fusion (HMAF) approach provides more accurate predictions. To do this, we improve predictive reliability by merging mappings of saliency employing three attention components with two inputs and one attention component with three inputs. Additionally, it efficiently predicts bottom-up saliency maps and effectively handles both global and local contrast levels. Furthermore, the suggested technique has the capability to effectively emphasize various top-down elements, including human faces, individuals in intricate backdrops, and items situated at varying distances from the camera.
The suggested HMAF-based solution was one of many that had the greatest computational difficulty estimated using information from the NCTU dataset. It requires around an hour to train the described HMAF on a GPU supported by NVIDIA TITAN V and a CPU with a 3.0 GHz capability from Intel i5-8500. On a 640 by 480 photo, deduction utilizing the planned HMAF requires approximately 0.01 second. Our approach possesses an ultimately minimal computational burden, rendering it suitable for implementation in real-time image reformat/creation systems.

5. Conclusion

The above discussion presented of the currently done research procedures on image detection approaches utilizing stereoscopic videos provide a clear indication of better efficacy of saliency integrated fusion based approaches to be more effective in terms of depth estimation and error correction.
The reason, as found from the experimental results is due to fusion strategy that merges the potency and computational power of unimodules to optimize the image detection model. Saliency criteria also plays as an important tool in feature extraction and attention based estimation of depth information. Overall, it can be concluded thata multimodal adaptive fusion network with saliency features, built on the principle of attention based extraction of three-dimensional data achieves good results.
Certainly, the model without saliency features showed much slower and faulty outputs than the saliency fitted multimodal adaptive fusion model of image identification from 3D spectroscopic videos.
Comparatively, the DNN model of spectroscopy based image detection that employed depth map strategy without saliency method gave errors and is slower in its detection process. The dual input attention module takes into account the relative relevance of the RGB and depth modes. Furthermore, for the RGB-D significance prediction, the multimodule attention module give varying weights to the fusion significance features at various levels.

Thus, compared to single module approaches, the experimental findings demonstrate that the for doing image detection tasks by utilizing spectroscopic video integrated with saliency features, multimodule models show significance prediction method is the best.

6. References

1. Banitalebi-Dehkordi, A., Nasiopoulos, E., Pourazad, M. T., & Nasiopoulos, P. (2016). Benchmark three-dimensional eye-tracking dataset for visual saliency prediction on stereoscopic three-dimensional video. *Journal of Electronic Imaging*, *25*(1), 013008. https://doi.org/10.1117/1.JEI.25.1.013008
2. Banitalebi-Dehkordi, A., & Nasiopoulos, P. (2018). Saliency inspired quality assessment of stereoscopic 3D Video. *Multimedia Tools and Applications*, *77*(19), 26055–26082. https://doi.org/10.1007/s11042-018-5837-4
3. Chacko, L., S, P., & Jayakumar, A. (2016). Salient Region Extraction for 3D-Stereoscopic Images. *International Journal of Engineering Research and General Science*, *4*(3).
4.] C.Chamaret, S.Godeffroy, P. Lopez, andO. LeMeur, (2017) "Adaptive3d rendering based on region-of-interest," *inProceedings of SPIE*, vol. 7524,2010,p.75240V.
5. Chen, Y., Pan, Y., Song, M., & Wang, M. (2015). Image retargeting with a 3D saliency model. *Signal Processing*, *112*, 53–63. https://doi.org/10.1016/j.sigpro.2014.11.001
6. Chen, H., Li, Y., Su, D., (2019). Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection. *Pattern Recognition* 86, 376–385. https://doi.org/10.1016/j.patcog.2018.08.007
7. Cheng, H., Zhang, J., An, P., Liu, Z., (2015). A Novel Saliency Model for Stereoscopic Images, in 2015 *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, Adelaide, Australia, pp. 1–7. https://doi.org/10.1109/DICTA.2015.7371220
8. Cheng, H., Zhang, J., Wu, Q., An, P., & Liu, Z. (2017). Stereoscopic visual saliency prediction based on stereo contrast and stereo focus. *EURASIP Journal on Image and Video Processing*, *2017*(1), 61. https://doi.org/10.1186/s13640-017-0210-5
9. E. Potapova, M. Zillich, and M. Vincze, (2011) Learning what matters: combining probabilistic models of 2d and 3d saliency cues, *Computer Vision Systems*, pp. 132-142,
10. Fang, Y., Wang, J., Narwaria, M., Le Callet, P., & Lin, W. (2013). Saliency detection for stereoscopic images. *2013 Visual Communications and Image Processing (VCIP)*, 1–6. https://doi.org/10.1109/VCIP.2013.6706346
11. Fang, Y., Wang, J., Yuan, Y., Lei, J., Lin, W., & Callet, P. L. (2016). Saliency-based stereoscopic image retargeting. *Information Sciences*, *372*, 347–358. https://doi.org/10.1016/j.ins.2016.08.062
12. Fang, Y., Ding, G., Li, J., Fang, Z., (2019a). Deep3DSaliency: Deep Stereoscopic Video Saliency Detection Model by 3D Convolutional Networks. *IEEE Trans. on Image Process*. 28, 2305–2318. https://doi.org/10.1109/TIP.2018.2885229
13. Fang, Y., Lei, J., Li, J., Xu, L., Lin, W., Callet, P.L., (2017a). Learning visual saliency from human fixations for stereoscopic images. *Neurocomputing* 266, 284–292. https://doi.org/10.1016/j.neucom.2017.05.050
14. Fang, Y., Lin, W., Fang, Z., Lei, J., Le Callet, P., Yuan, F., (2014a). Learning visual saliency for stereoscopic images, in 2014 *IEEE International Conference on Multimedia and Expo Workshops (ICMEW). IEEE*, Chengdu, China, pp. 1–6. https://doi.org/10.1109/ICMEW.2014.6890709
15. Fang, Y., Wang, J., Narwaria, M., Callet, P.L., Lin, W., (2014b). Saliency Detection for

Stereoscopic Images. *IEEE Transactions On Image Processing* 23.
16. Fang, Y., Wang, J., Narwaria, M., Le Callet, P., Lin, W., (2014c). Saliency Detection for Stereoscopic Images. *IEEE Trans. on Image Process*. 23, 2625–2636. https://doi.org/10.1109/TIP.2014.2305100
17. Fang, Y., Wang, J., Narwaria, M., Le Callet, P., Lin, W., (2014d). Saliency Detection for Stereoscopic Images. *IEEE Trans. on Image Process*. 23, 2625–2636. https://doi.org/10.1109/TIP.2014.2305100
18. Fang, Y., Wang, J., Narwaria, M., Le Callet, P., Lin, W., (2014e). Saliency Detection for Stereoscopic Images. *IEEE Trans. on Image Process*. 23, 2625–2636. https://doi.org/10.1109/TIP.2014.2305100
19. Fang, Y., Wang, J., Yuan, Y., Lei, J., Lin, W., Callet, P.L., (2016). Saliency-based stereoscopic image retargeting. *Information Sciences* 372, 347–358. https://doi.org/10.1016/j.ins.2016.08.062
20. Fang, Y., Zhang, C., Huang, H., Lei, J., (2019b). Visual Attention Prediction for Stereoscopic Video by Multi-Module Fully Convolutional Network. *IEEE Trans. on Image Process*. 28, 5253–5265. https://doi.org/10.1109/TIP.2019.2916766
21. Fang, Y., Zhang, C., Li, J., Lei, J., Perreira Da Silva, M., Le Callet, P., (2017b). Visual Attention Modeling for Stereoscopic Video: A Benchmark and Computational Model. *IEEE Trans. on Image Process.* 26, 4684–4696. https://doi.org/10.1109/TIP.2017.2721112
22. Hachaj, T. (2023). Adaptable 2D to 3D Stereo Vision Image Conversion Based on a Deep Convolutional Neural Network and Fast Inpaint Algorithm. *Entropy*, *25*(8), 1212. https://doi.org/10.3390/e25081212
23. Huang, N., Liu, Y., Zhang, Q., & Han, J. (2021). Joint Cross-Modal and Unimodal Features for RGB-D Salient Object Detection. *IEEE Transactions on Multimedia*, *23*, 2428–2441. https://doi.org/10.1109/TMM.2020.3011327
24. Lagendijk, R., Franich, R. E., & Hendriks, E. (2009). Stereoscopic Image Processing. In V. Madisetti, *Video, Speech, and Audio Signal Processing and Associated Standards* (Vol. 20096073, pp. 1–11). CRC Press. https://doi.org/10.1201/9781420046090-c22
25. Lv, Y., & Zhou, W. (2020). Hierarchical Multimodal Adaptive Fusion (HMAF) Network for Prediction of RGB-D Saliency. *Computational Intelligence and Neuroscience*, *2020*, 1–9. https://doi.org/10.1155/2020/8841681
26. Sahu, G., & Vechtomova, O. (2021). Adaptive Fusion Techniques for Multimodal Data. *Conference of the European Chapter of the Association for Computationa lLinguistics*, 3156–3266.
27. Wang, J., Fang, Y., Narwaria, M., Lin, W., Le Callet, P., (2014). Stereoscopic image retargeting based on 3D saliency detection, in 2014 IEEE *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Florence, Italy, pp. 669–673. https://doi.org/10.1109/ICASSP.2014.6853680
28. Wang, J., Silva, M.P.D., Callet, P.L., Ricordel, V., (2013). A computational model of stereoscopic 3D visual saliency.
29. Wang, W., Shen, J., Yu, Y., Ma, K.-L., (2017). Stereoscopic Thumbnail Creation via Efficient Stereo Saliency Detection. IEEE Trans. Visual. *Comput. Graphics* 23, 2014–2027. https://doi.org/10.1109/TVCG.2016.2600594
30. Wang, X., Ma, L., Kwong, S., Zhou, Y., (2018). Quaternion representation-based visual saliency for stereoscopic image quality assessment. *Signal Processing* 145, 202–213. https://doi.org/10.1016/j.sigpro.2017.12.002

31. Wang, S., Zhou, Z., Jin, W., & Qu, H. (2018). Visual saliency detection for RGB-D images under a Bayesian framework. *IPSJ Transactions on Computer Vision and Applications*, *10*(1), 1. https://doi.org/10.1186/s41074-017-0037-0
32. Wang, X., Ma, L., Kwong, S., & Zhou, Y. (2018). Quaternion representation based visual saliency for stereoscopic image quality assessment. *Signal Processing*, *145*, 202–213. https://doi.org/10.1016/j.sigpro.2017.12.002
33. Zhang, P., Liu, J., Wang, X., Pu, T., Fei, C., Guo, Z., (2020). Stereoscopic video saliency detection based on spatiotemporal correlation and depth confidence optimisation. *Neurocomputing* 377, 256–268. https://doi.org/10.1016/j.neucom. (2019).10.024
34. Zhang, Q., Wang, X., Wang, S., Li, S., Kwong, S., Jiang, J., (2019). Learning to Explore Intrinsic Saliency for Stereoscopic Video, in 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Long Beach, CA, USA, pp. 9741–9750. https://doi.org/10.1109/CVPR.2019.00998
35. Zhang, Q., Wang, X., Wang, S., Sun, Z., Kwong, S., Jiang, J., (2020). Learning to Explore Saliency for Stereoscopic Videos Via Component-Based Interaction. *IEEE Trans. on Image Process*. 29, 5722–5736. https://doi.org/10.1109/TIP.2020.2985531
36. Zhang, Q., Wang, X., Wang, S., Sun, Z., Kwong, S., & Jiang, J. (2020). Learning to Explore Saliency for Stereoscopic Videos Via Component-Based Interaction. *IEEE Transactions on Image Processing*, *29*, 5722–5736. https://doi.org/10.1109/TIP.2020.2985531