Article

Holoscopic Elemental Image-Based Disparity Estimation using Multi-scale Multi-window Semi-Global Block Matching

Bodor Almatrouk , **Hongying Meng** , **Mohammad Rafiq Swash**

Department of Electronic and Electrical Engineering, Brunel University London, London; bodor.almatrouk@brunel.ac.uk, hongying.meng@brunel.ac.uk, rafiq.swash@brunel.ac.uk
* Correspondence: hongying.meng@brunel.ac.uk

Abstract: Holoscopic imaging, which a single aperture to acquire full-colour spatial images like the fly's eye by gently altering angles between nearby lenses with a micro-lens array. Due to its simple data collection and visualisation method, which provides robust and scalable spatial information, and motion parallax, binocular disparity, and convergence, this technique may be able to overcome traditional 2D imaging issues like depth, scalability, and multi-perspective problems. A novel disparity map-generating method uses angular information in a single Holoscopic image's micro-images, or Elemental Images (EI), to create a scene's disparity map. Not much research has used EIs instead of Viewpoint Images VPIs for disparity estimation. This study investigates whether angular perspective data may replace spatial orthographic data. Using noise reduction and contrast enhancement, EIs with low resolution and texture are pre-processed to calculate the disparity. The Semi-Global Block Matching (SGBM) technique is used to calculate the disparity between EIs pixels. A multi-resolution approach overcomes EIs' resolution constraints, and a content-aware analysis dynamically modifies the SGBM window size settings to generate disparities across different texture and complexity levels. A background mask and nearby EIs with accurate backgrounds detect and rectify EIs with erroneous backgrounds. Our method generated disparity maps that outperformed two state-of-the-art deep learning algorithms and VPIs in real images.

Keywords: Holoscopic; Elemental Images; Viewpoint Images, Micro-lenses, Disparity, SGBM

1. Introduction

Depth estimation from Holoscopic images is a promising technique that has gained interest recently due to its advantage of calculating depth using a single-aperture camera. Holoscopic cameras are based on the same fundamental principles as conventional cameras but with an additional array of micro-lenses (MLA) in front of the image sensor. In traditional cameras, the main lens translates the object plane into the camera's image plane. The micro-lenses focus light beams from various directions onto a single pixel, thereby capturing the scene in three dimensions.

The pixels behind each micro-lens record the same data as traditional cameras but with greater precision by measuring information from different angles, as shown in Fig. 1. The images formed behind each micro-lens, known as the Elemental Images (EIs), represent unique angles of light incidence. Thus, by analysing the EIs, the location and orientation of each light beam can be determined on a pixel-by-pixel basis. A sub-aperture image of a scene, or a Viewpoint Image (VPI), is created by re-sampling pixels from the same locations across the EIs. The EIs provide angular information, whereas VPIs provide spatial information.

Citation: Almatrouk B, Meng H and Swash R, Holoscopic Elemental Image-Based Disparity Estimation using Multi-scale Multi-window Semi-Global Block Matching. *Journal Not Specified* **2024**, *1*, 0. https://doi.org/

Received:
Revised:
Accepted:
Published:

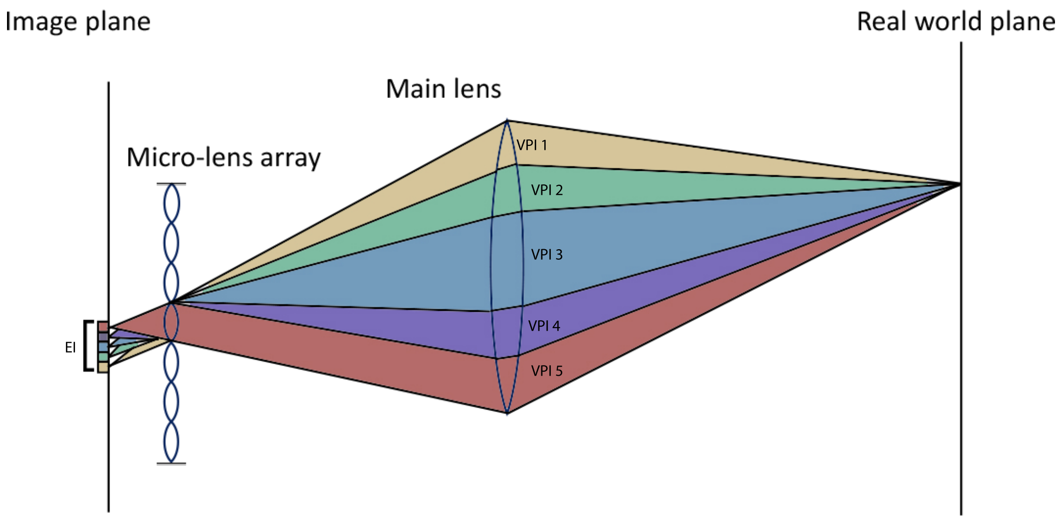


Figure 1. Light beams from various perspectives (VPIs) hitting the same EI in the Holoscopic sensor.

Traditionally, disparity estimation is performed on VPIs, which encompass the entire scene from a certain perspective, whereas EIs only include a portion of it. VPIs share visual characteristics with 2D orthographic stereo images, allowing existing stereo image-based disparity estimation methods to be applied with few adjustments. Additionally, VPIs can be up-sampled using shift and integration methods [1].

However, extracting VPIs requires mapping the information gathered on the sensor to reconstruct the scene, which is not always straightforward and can sometimes lead to strong aliasing artefacts [2–4]. Lens error correction and camera calibration must be performed initially, as depicted in Fig. 2 (a) and (b), showing extraction with and without lens correction respectively [5]. The geometry of the scene must also be considered during VPI creation to avoid image artefacts in areas not 'in focus' [6]. Additionally, some micro-lens array designs feature multiple micro-lens sizes with different focal lengths, making it impractical to extract pixels from the same location across all EIs as seen in Fig. 2 (c) and (d). The convergence of light rays from multiple VPIs might result in overlapping on the image sensor, complicating the separation and extraction of individual rays. Therefore, selecting a 'patch' of pixels from each EI might be more effective in increasing the resolution and reducing the artefacts. Yet, it is more challenging than choosing a single pixel as these patches depend on the depth level within the scene; thus, using the same patch size throughout the entire scene could result in a distorted VPI. For instance, the ideal patch size for displaying the foreground can be excessively large for the background, leading to the occurrence of artefacts in the background.

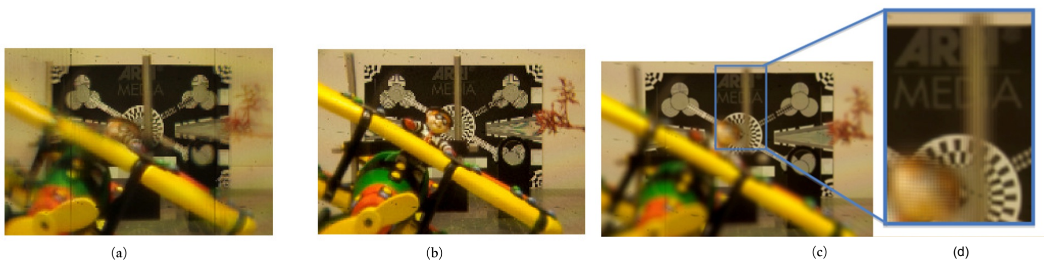
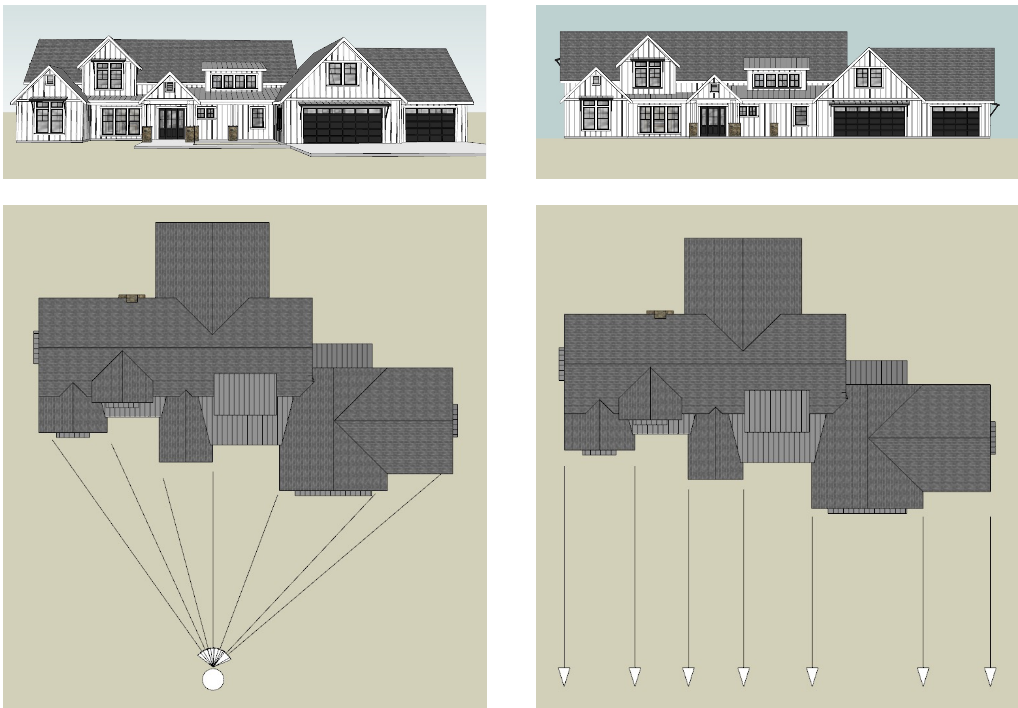


Figure 2. VPI (25, 25) extracted (a) without lens distortion correction. (b) with lens distortion correction. (c) Shows image artefacts in the foreground. (d) Close-up view: artefacts in the foreground with background corrected. [5]

Extracting VPIs from Holoscopic images is time-consuming and requires significant storage due to the large number of VPIs generated. Estimating depth from video frames or in real-time scenarios is particularly challenging due to the large number of frames [7]. For these reasons, EIs provide a more straightforward method for estimating disparities, requiring only lens correction as pre-processing. In this paper, disparity estimation using perspective EIs is employed, contrary to conventional methods that use extracted, corrected, and up-sampled VPIs.

Perspective and orthographic projection are two types of 3D projection. As seen in Fig. 3, perspective projection is comparable to the human visual system in which parallel lines in an image appear to converge at a single point; the closer the object is to the point of convergence, the smaller it appears (change in scale). The orthographic (orthogonal to the scene) projection assumes parallel lines will continue to be parallel and disregards the scaling impact.

Understanding depth via a perspective projection is far more precise than using an orthographic technique because, in perspective depth, every light ray is tracked to the precise pixel of its source, unlike in orthographic depth, where light is considered to be emanating from infinity [8]. Although perspective projection has shown more accurate disparity estimation results [9–11], most depth estimation algorithms are performed on orthographic images due to the simplicity of the capturing mechanisms.



(a) (b)

Figure 3. (a) Perspective Projection. (b) Orthographic Projection

As seen in Fig. 1, the EIs in the Holoscopic setup record light from different angles, resulting in perspective images that contain angular information. Conversely, VPIs are obtained from various locations on the primary lens, replicating different viewpoints. These images typically exhibit orthographic projection, predominantly capturing surface characteristics. The differentiation here between EIs and VPIs is linked to their ways of spatial representation [12].

2. Methodology

A single Holoscopic image records the scene's spatial and angular details. Hence, it is possible to compute the scene's depth map from a single shot. Our proposed method as seen in Fig. 4 begins a Pre-processing is carried out on the EIs, which is crucial before computing the disparity to improve their quality, as they inherently have low resolution

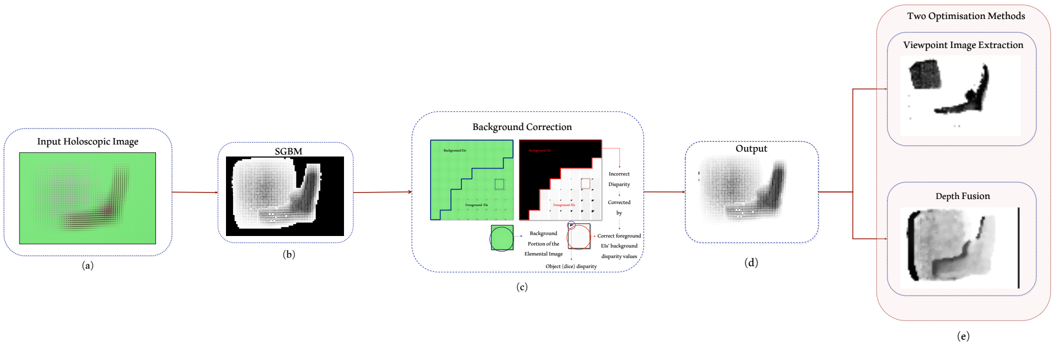


Figure 4. The disparity estimation from EIs pipeline. (a) Input raw Holoscopic image. (b) Disparity map using Multi-Resolution Content-Aware SGBM matching. (c) Background correction using background/ foreground mask. (d) Output Disparity image after SGBM and background correction. (e) Output: two optimisation results. Top: extracted central VPI. Bottom: fusing depth from multiple EIs

and lack texture. This procedure consists of two primary stages: noise reduction by bilateral filtering and contrast enhancement via histogram equalisation.

The disparity among EI pixels is computed via the Semi-Global Block Matching (SGBM) algorithm [13], which is favoured due to its flexibility to adapt to the unique features of EIs and its optimal balance between precision and computing efficiency. The SGBM algorithm is enhanced through a multi-resolution approach to address the limitations of EIs in terms of resolution. This involves creating an upscaled pyramid of EIs to capture details at different scales and performing a content-aware analysis to adaptively adjust the SGBM window size parameters. This ensures optimal estimation of disparities across various texture and complexity levels within the EIs. Ultimately, a weighted least squares (WLS) filter is employed to further enhance the optimisation process.

EIs are known to be low in resolution, lack texture, and only capture a portion of the scene. Several deep-learning models have been designed to estimate disparity maps, including many specifically designed for VPIs. Deep learning necessitates a substantial and comprehensive dataset specifically designed for Holoscopic imagery. Pre-existing, pre-trained deep learning stereo-matching solutions would not be compatible with EIs due to differences in training data properties. These solutions are mostly learned using high-quality images, while EIs have low resolution and lack texture. Deep learning algorithms have the potential to be highly effective in stereo vision tasks, but their effectiveness is contingent upon the quality and range of the training data. If the training data lacks sufficient representation of scenarios including low resolution, limited texture, and narrow disparity ranges, the model may exhibit poor generalisation in these settings. Deep learning algorithms may encounter difficulties in generating intricate details in such situations, resulting in unclear outcomes.

2.1. Pre-Processing of Elemental Images

Before initiating a disparity estimate on the EIs, it is crucial to carry out pre-processing on the EIs to adequately prepare them to achieve an improved outcome. EIs exhibit low resolution and limited texture. Therefore, while implementing pre-processing techniques, it is crucial to eliminate noise while preserving the critical features.

2.1.1. Noise Reduction through Bilateral Filtering

Applying image blurring is a conventional technique for diminishing image noise. Particularly with images that have minimal texture, such as the background, there may be instances where a stepping effect occurs. This effect is caused by discontinuous disparity levels, resulting in noticeable "steps" in areas with reduced changes in depth as seen in the textured map in Fig. 5. The limited resolution, subtle variations in lighting, limited bit depth, and lack of texture can cause seamless transitions to look like sudden shifts. However, using image blurring will inevitably cause a loss of fine information such as edges, hence reducing the accuracy of disparity estimation. To address this issue, the application of bilateral filtering [14] is employed. This technique, known for its ability to preserve edges, is considered an advanced way of blurring.

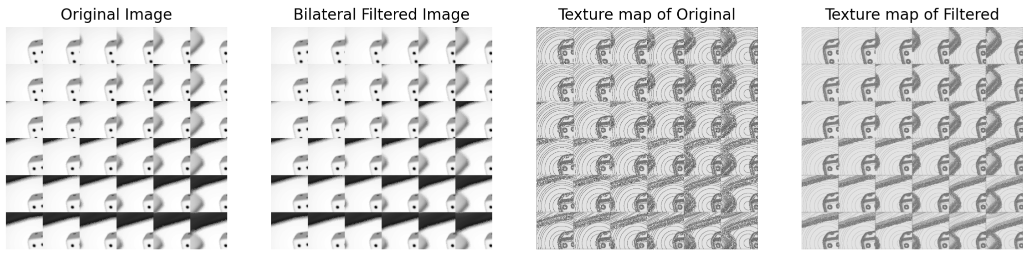


Figure 5. EI before and after applying bilateral filtering. As seen in the texture map of the original image, there is a noticeable stepping effect in the background. Although the filter did not eliminate it, it did assist reduce the impact while maintaining edge information.

Bilateral blurring is applied to each EI to reduce the noise:

$$I_{\text{filtered}}\,(p) = \frac{1}{W_p} \sum_{q \in S} I(q) \cdot f_r(\|I(p) - I(q)\|) \cdot f_s(\|p - q\|) \tag{1}$$

Let $I_{\text{filtered}}\,(p)$ represent the filtered intensity of pixel p, $I(q)$ denote the intensity of the next pixel, S be the set of pixels surrounding p, and W be the normalised factor. The variable f_r represents the spatial range of the kernel, which corresponds to the dimensions of the neighbouring region. On the other hand, f_s denotes the minimum magnitude required for an edge to be detected. This procedure ensures that only pixels with similar intensity levels to the core pixel are considered for blurring while maintaining distinct intensity fluctuations. A lower value of f_r leads to a more distinct edge. As the value of f_s tends towards infinity, the equation approaches convergence to a Gaussian blur.

2.1.2. Contrast Enhancement via Histogram Equalisation

Due to their low resolution and the settings under which they are captured (tiny micro-lenses), EIs often experience a lack of contrast. Histogram equalisation is commonly employed to enhance image contrast by spreading the intensity levels, hence boosting feature visibility by:

$$I_{\text{equalised}}\,(p) = H(I(p)) \tag{2}$$

where $I(p)$ is pixel p original intensity, $I_{\text{equalised}}\,(p)$ represent the equalised intensity of pixel and H is the is the cumulative distribution function.

2.2. Content-aware Multi-resolution Disparity Estimation using Semi-Global Block Matching

2.2.1. Overview

Deriving disparity from EIs using SGBM presents challenges, mostly attributed to the presence of low texture and low resolution. Upsampling the EIs would lead to data loss, resulting in the introduction of noise and a decrease in image quality. Many studies [15,16] have investigated the computation of disparity at various resolutions to enhance the accuracy of disparity maps, particularly in the context of developing deep learning models. While rescaling EIs may result in a loss of image quality, calculating the disparity at multiple resolutions instead of merely one upsampled resolution still is an effective approach for handling varying levels of details and textures. Lower resolutions may result in the loss of some details in the scene, while higher resolutions may exhibit an inconsistent overall structure. This indicates a trade-off between maintaining structural consistency and capturing high-frequency details based on the input resolution [16].

Content information can vary across different regions within EIs, particularly at varying resolutions. By utilising a content-aware approach, the disparity window size can be dynamically modified according to the characteristics of each location, resulting in more accurate disparity estimations. When working with areas that have a high level of texture, using a smaller window size would be advantageous in capturing intricate details. Conversely, in areas that lack texture, a larger window size can be used to minimise noise.

2.2.2. Multi-resolution Elemental Images Pyramid

Typically, when constructing a pyramid with multiple resolutions for any objective, the procedure commences by taking the original image and reducing its size. However, when it comes to EIs, the images are already of low resolution. Creating a pyramid by progressively down-sampling them will result in extremely small images that lack significant information. Thus, in the instance of the EIs, the pyramid is formed by enlarging the EIs into 2 additional layers and downsampling the image by one layer as seen in Fig. 6, enabling the algorithm to encompass characteristics that span from large-scale structures at lower levels to intricate details at higher resolutions. Starting with the EI of the original size as the base level L_0. With each level increased by a factor of 2 using bicubic interpolation [17]. A minimum resolution threshold is implemented to prevent further down-sampling of EIs with extremely low resolution. If the value of EI is less than 40×40, the down-sampling step is omitted.

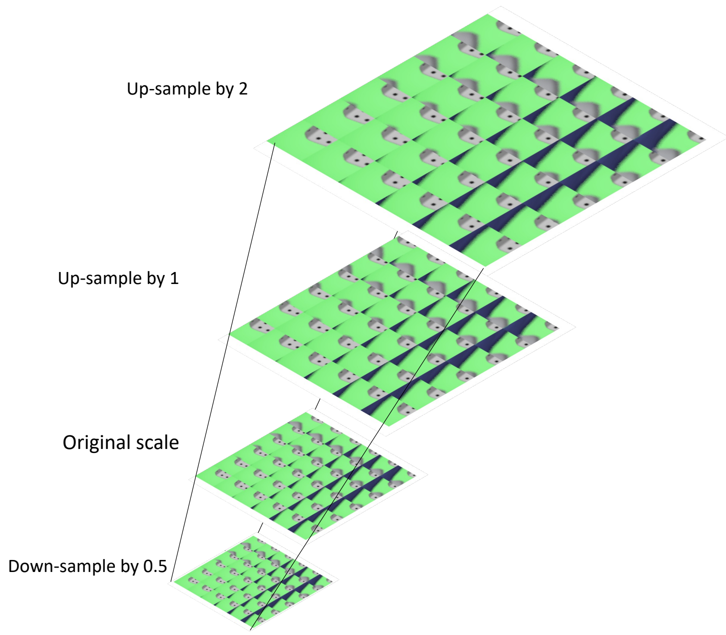


Figure 6. Multi-resolution pyramid of EIs

2.2.3. Multi-Resolution Content Analysis

Content-aware analysis is an essential process for evaluating the visual attributes in the EIs. Its purpose is to optimise the window size parameters used in disparity estimation based on the complexity and textures present at different scales. This analysis is particularly valuable for adjusting window size parameters at both single and multiple scales to enhance the precision and resilience of the disparity.

EIs possess a high degree of sensitivity. Consequently, a simpler approach involving edge segmentation and texture analysis is employed. The Sobel filter is utilised for accomplishing edge detection. The filter's sensitivity is contingent upon the resolution of the images. Low-resolution images necessitate a higher threshold for detecting significant

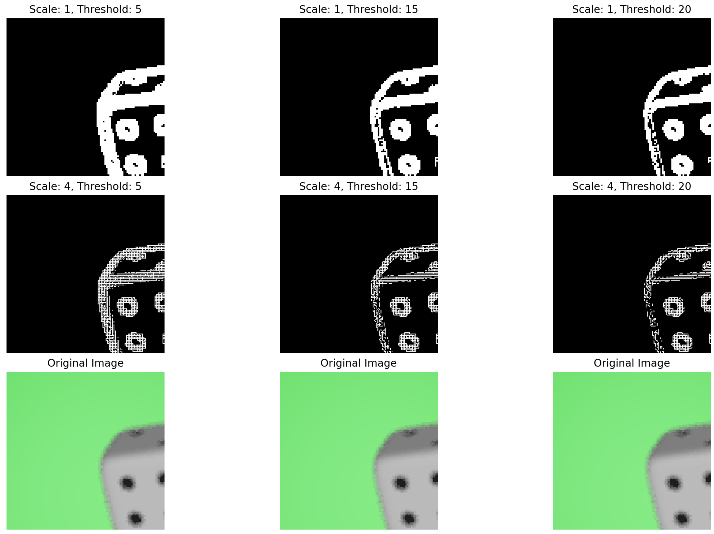


Figure 7. Examples of extreme edge thresholds show that the sensitivity of the filter depends on the resolution of the image. High-resolution images need a lower threshold to identify finer structures, while low-resolution images need a higher threshold to identify significant features.

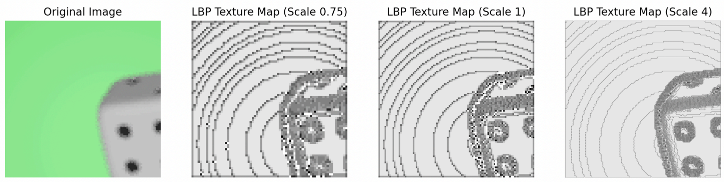


Figure 8. LBP texture maps across different scales before filtering to show the effect

structures, while high-resolution images require a lower threshold to identify finer details as depicted in Fig. 7.

Textures are ideal to identify intensity patterns which is great for identifying regions for disparity estimation. Local Binary Patterns (LBP) are used in this case to identify the textures in the EIs. Here, the focus is on larger patterns at lower resolutions and finer textural details at higher resolutions.

$$LBP_n(p) = \sum_{k=0}^{P-1} 2^k \cdot \mathbf{1}(I_n(p_k) \geq I_n(p)) \tag{3}$$

where LBP is computed for pixel p located at location n within the image used to classify the texture. P represents the total number of pixels neighbouring to p, with the summation ranging from $k = 0$ to $P - 1$. 2^k represents the weighting factor assigned to each neighbouring element, which is determined by its location. The neighbouring pixel's (p_k) intensity is compared with the central pixel $I_n(p)$. $\mathbf{1}(I_n(p_k) \geq I_n(p))$ returns 1 if it is true and 0 if false. Texture maps across different scales are shown in Fig. 8.

To enhance simplicity and preserve time, the edge map E and texture map T are combined to then form a dynamic adaptive window for the computation of disparities.

$$F(x, y) = \alpha \cdot E(x, y) + (1 - \alpha) \cdot T(x, y) \tag{4}$$

The combined feature at pixel (x, y) is denoted as F, and it is influenced by a weighted factor, α, which ranges from 0 to 1. This factor determines the appropriate ratio between edge and texture data. The value of α has been cautiously adjusted to achieve the ideal result for each image.

2.2.4. Multi-Resolution Multi-Window Disparity Estimation using SGBM

Dynamic Window: Given that the EI's resolution varies from 50x50 to around 400x400, it is necessary to select a range of window sizes. The value of W_{min} is selected to be around 5% of the minimum resolution, resulting in an amount of 5. Similarly, the value of W_{max} is chosen to be roughly 20% of the resolution, resulting in a value of 80. The size of the window adjusts according to the value of the feature map F. The dynamic window size W at pixel (x, y) can be calculated by:

$$W(i,j) = W_{\text{min}} + (W_{\text{max}} - W_{\text{min}}) \cdot (1 - F_{\text{norm}}(i,j)) \tag{5}$$

Greater values of F_{norm}, representing the normalised F values, will result in the selection of smaller windows for more complicated regions, and vice versa.

Semi-Global Block Matching Disparity: The disparity is calculated by comparing blocks of pixels along the epipolar line and obtaining the associated vertical displacement, as demonstrated in our previous work [18]. This problem can be represented by a comprehensive cost function:

$$E(D) = \sum_{d \in D} \left(C(d) + \sum_{d' \in N(d)} P_1 I_{\{|d-d'|=1\}} + \sum_{d'' \in N(d)} P_2 I_{\{|d-d''|>1\}} \right) \tag{6}$$

where I is a function that indicates whether an input is true or false and returns 1 or 0 accordingly. (d) is the chosen disparity's data term similarity metric. A 3D cost structure is used to hold each similarity cost, and this process is repeated for each pixel block, with a cost of d. The 3D structure stack's minimal costs stand for possible disparity estimates [18].

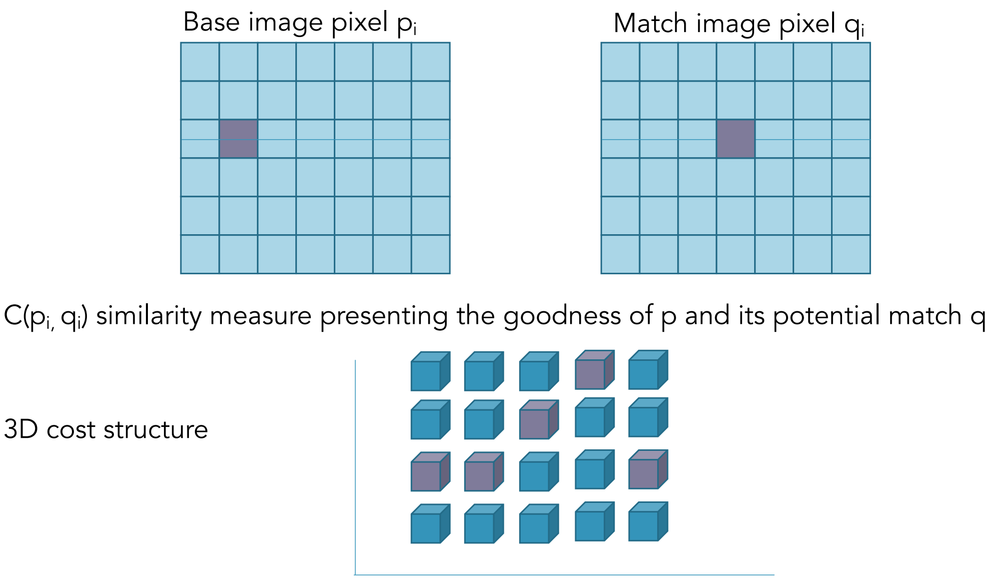


Figure 9. The produced minimal costs are not highly distinctive, which could result in incorrect disparity estimation [18].

Disparity Aggregation: The resulting minimum costs may lack significant distinctiveness, thus resulting in an incorrect assessment of disaprity. This issue is addressed by employing cost aggregation within these 3D cost structures. The total cost is determined by aggregating the lowest costs across various image paths. A total of eight paths were utilised in this paper [18]. Potential cost values are pooled, and a weighted summing of these cost possibilities is conducted. The weights are obtained from the normalised feature map $F_{\text{norm}}(x, y)$, which characterises the contents (texture and edges) at each scale level. The feature map undergoes normalisation:

$$F_{\text{norm}}(x,y) = \frac{F(x,y)}{\sum_L F_L(x,y)} \tag{7}$$

The function F_{norm} represents the normalised feature for pixels (x,y), whereas L is the scale level. Normalising the feature map guarantees that, throughout the content-aware analysis, disparities from all resolutions contribute proportionally. Thus, the final content-aware disparity map D_{final} can be represented as:

$$D_{\text{final}}(x,y) = \sum_i \left(\frac{F_i(x,y)}{\sum_j F_j(x,y)} \right) \cdot D_L(x,y) \tag{8}$$

where D_L is the disparity optimised at each level of resolution. To achieve greater accuracy, a higher weight is assigned to the original scale since this method is still sensitive to multiple scales:

$$\begin{aligned} D_{\text{final}}(x,y) = & \left(\frac{\alpha \cdot F_{L_0}(x,y)}{\sum_j F_j(x,y) + (\alpha - 1) \cdot F_{L_0}(x,y)} \right) \cdot D_{L_0}(x,y) \\ & + \sum_{i \neq L_0} \left(\frac{F_i(x,y)}{\sum_j F_j(x,y) + (\alpha - 1) \cdot F_{L_0}(x,y)} \right) \cdot D_L(x,y) \end{aligned} \tag{9}$$

The expression $\sum_j F_j(x,y) + (\alpha - 1) \cdot F_{L_0}(x,y)$ ensures normalisation for assigning a larger weight to F_{L_0}, where F_{L_0} represents the feature map at the original level and α is the weighting factor.

Penalty terms P_1 and P_2 are introduced, which are based on the difference in neighbourhood disparities, where $N(d)$ is the neighbour of d. Accordingly, for each pixel, all its neighbouring pixels along the routes are analysed; the greater the difference between the lateral parallax axis of the pixel and its neighbours, the greater the penalty, resulting in a considerable increase in the source value of the matching costs (Fig. 10). This procedure ensures a smooth surface by forcing the strings along the path to be somewhat continuous. This process is repeated for each path and each correspondence in the image to get the final cost.

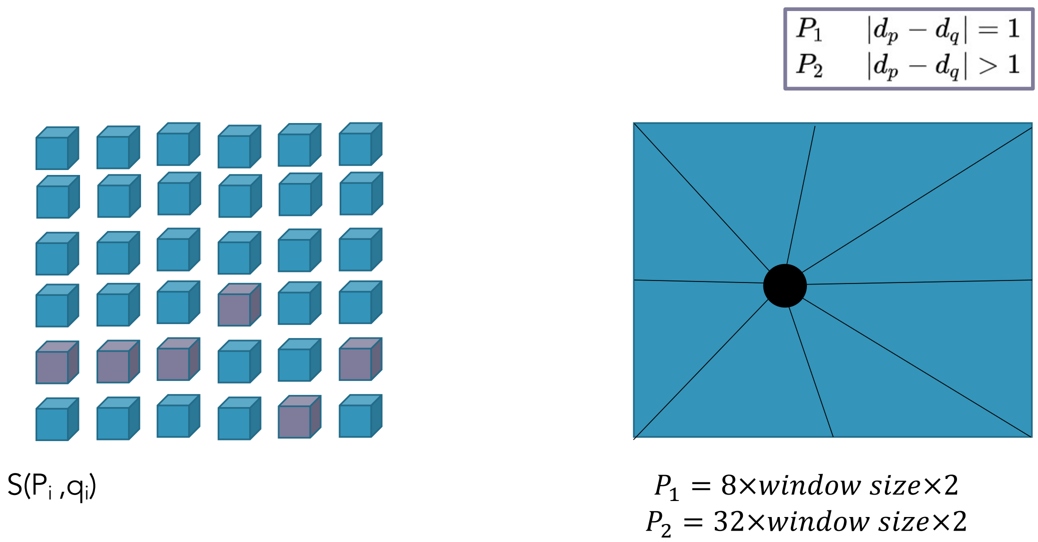
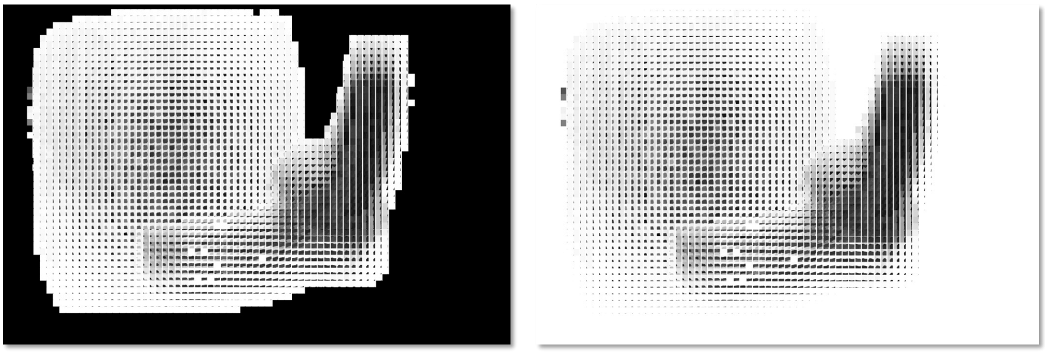


Figure 10. The ultimate cost is the sum of the least costs along picture routes. 8 pathways were used. Cost possibilities are pooled and weighted. P_1 and P_2 are based on neighbourhood disparities, where $N(d)$ is d's neighbour [18].

To minimise the noise in the computed disparity image, a weighted least squares (WLS) filter [19] is applied [18]. The WLS filter, a well-known edge-preserving smoothing technique, has weights that highly depend on the image gradients. The final disparity image can be seen in Fig. 11.



(a) (b)

Figure 11. (a) The final disparity image using EIs before background correction. (b) The final disparity image using EIs after background correction. In this disparity map, the darker the pixel, the closer it is to the camera for clarity.

2.3. Background's Disparity Correction

EI-based disparity estimation allows for the recovery of angular information. However, as can be seen in Fig. 11 (a), the incorrect disparity may emerge from large texture-less areas such as the background because the EIs only represent segments of the whole scene. Thus, a solution is implemented in which first background extraction is performed to create a background mask, and then the disparity is corrected.

Initially, a disparity map D of the same size as the Holoscopic image is filled with zeros. Then the EIs are iterated over in the Holoscopic to select the left and right pairs:

$$EI_L = EI(i,j), EI_R = EI(i,j+1);\text{ where } i \in [0,n], j \in [0,m) \tag{10}$$

where i and j are the EIs' location in the Holoscopic image of size (n,m). The disparity for each left and right pair of EI is computed and the resulting disparity is filtered. $D(i,j)$ is filled with the computed disparity map.

To separate background EIs from foreground EIs, the background threshold value bg_{th}, which is in the range [0, 1] based on the disparity map, is defined. The ratio between non-zero disparity values and the total number of values in the disparity map is computed r. If this ratio is greater than bg_{th}, the $EI(i,j)$ is labelled as a foreground EI, otherwise as a background EI. Increasing the value of bg_{th} will add more images to background EIs, and vice versa.

$$EI(i,j) \begin{cases} M_{bg}, r <= bg_{\text{th}} \\ M_{fr}, r > bg_{\text{th}} \end{cases} \tag{11}$$

Background EIs' disparity values are corrected using the correct background disparity values in the foreground EIs, as seen in Fig. 12 using colour descriptors [20,21]. To obtain a mask for background regions within foreground EIs (bgr_{fg}), the mean and standard deviation of each channel (RGB) of foreground EIs are generated. A pixel in the foreground of an EI is considered to be part of the background if its value is less than one standard deviation from the mean (across all three RGB channels). This presupposes that the majority of foreground image pixels are part of the background region. This implies that the average pixel value should be within 1 standard deviation of the intensity of the background pixels at the very least. Finally, calculate the mode of the disparity values for bgr_{fg}.

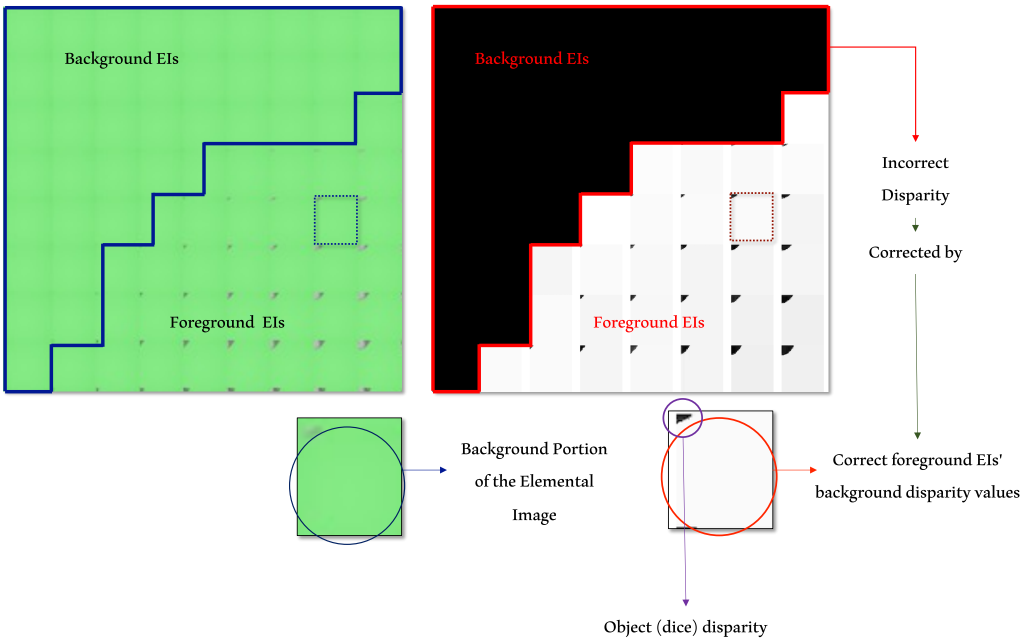


Figure 12. Background EIs' disparity values are corrected using the foreground EIs' background disparity values. Left: Holoscopic image showing background and foreground EIs. Right: Holoscopic disparity map showing incorrect background EIs being corrected by the background information of the correct disparity of the foreground EIs

Finally, the mode (mean or median) of the disparity values for bgr_{fg} is substituted for the disparity values (in D) of all background EIs. This acts as a disparity correction step for EIs that only contain background images since the stereo SGBM will fail to work for such pairs. Instead, the background disparity is corrected by replacing it with disparity from background regions in foreground EIs. The output result can be seen in Fig. 11 (b), where the background disparity information is fixed.

3. Evaluation

3.1. Dataset:

The methodology underwent three evaluations: one comparing the method on VPIs against EIs, another evaluating the method across multiple resolutions, and a third evaluating the method on the same dataset but against two other deep learning methods. The study utilised two Holoscopic datasets to determine the effectiveness of the methodology. The first is a synthetic dataset [22], that was specifically created to replicate the features of Brunel's Holoscopic full-frame camera sensor (Fig.13), which has a sensor size of 35 × 24 mm and a resolution of 40 megapixels, resulting in image dimensions of 7900 × 5300 pixels. The dataset has five EIs resolutions: 20 × 20, 40 × 40, 60 × 60, 80 × 80, and 100 × 100 pixels. The simulated images were used to evaluate different resolutions and compare them with deep learning techniques. The second dataset was acquired using Brunel's Holoscopic camera. This dataset is utilised because the synthetic one provides flawless VPI and EI pixel mapping, resulting in perfect VPIs that are free from lens effects, distortion, and artefacts. Hence, it is not feasible to directly compare the disparity outcomes between EIs and VPIs derived from the synthetic images.

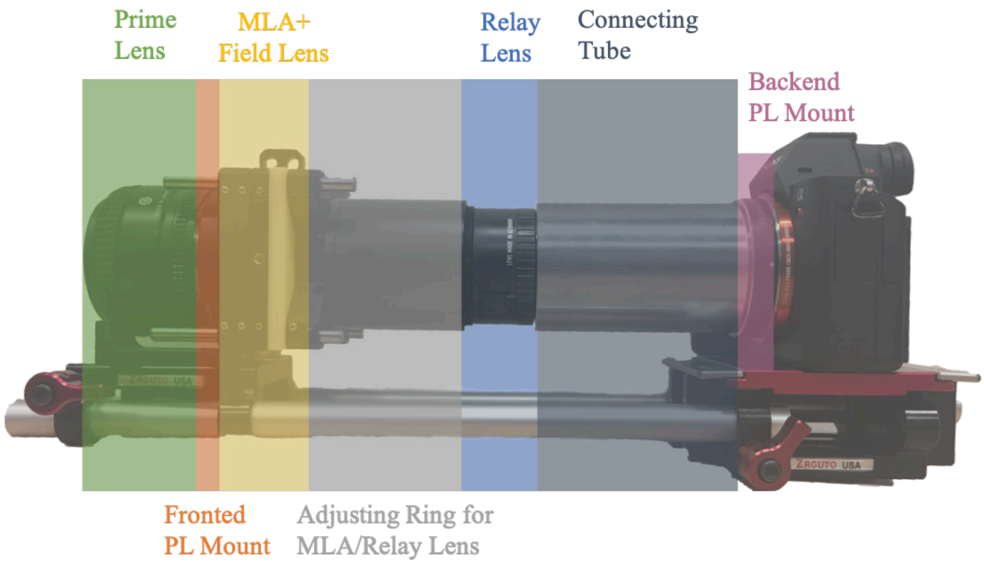


Figure 13. Brunel Holoscopic camera that includes a prime lens, a microlens array, a relay lens to focus light beams onto the sensor, and a CMOS imaging sensor.

3.2. Metrics:

Two types of metrics were used to assess the accuracy of the disparity estimation methodology: non-ground-truth metrics and ground-truth metrics. Ground truth-based metrics provide dependable evaluation outcomes, but real images from the Brunel camera lack ground truth disparity, necessitating alternative measurements.

Non-ground-truth Metrics: The consistency check metric, or left-right disparity consistency, evaluates disparity uniformity between left and right images, ensuring pixel correspondence. It's used for refining disparities by scanning both disparities to identify errors at the pixel level, with the error value indicating precision in the disparity map:

$$E = |\ d_{\mathrm{l}}(x,y) - d_{\mathrm{r}}(x - d_{\mathrm{l}}(x,y), y)\ | \leq \theta \tag{12}$$

The average error, E_{avg}, calculates the mean disparity error for each pixel:

$$E_{\mathrm{avg}} = \frac{1}{N} \sum_{x,y} |d_{\mathrm{l}}(x,y) - d_{\mathrm{r}}(x - d_{\mathrm{l}}(x,y), y)| \tag{13}$$

Edge alignment evaluates disparity near edges using the Sobel operator for edge detection. The Mean Absolute Error (MAE) and its normalised version assess disparity accuracy:

$$\mathrm{MAE} = \frac{1}{N} \sum_{(x,y)} |I_e(x,y) - d_e(x,y)| \tag{14}$$

$$\mathrm{MAE}_{\mathrm{norm}} = 1 - \frac{\mathrm{MAE}}{\mathrm{MAE}_{\mathrm{max}}} \tag{15}$$

Non-ground-truth metrics, though less reliable, provide insight into disparity errors.

Ground-truth Metrics: For synthetic datasets, ground-truth metrics include the Mean Absolute Error (MAE) for the average absolute difference between predicted and actual disparities, and the Percentage of Bad Pixels (PBP) for recognising significantly incorrect disparity pixels:

$$\mathrm{MAE} = \frac{1}{N} \sum_{x=1}^{W} \sum_{y=1}^{H} |d_{\mathrm{e}}(x,y) - d_{\mathrm{gr}}(x,y)| \tag{16}$$

$$PBP = \frac{1}{N_{\mathrm{P}}} \sum_{(x,y)} (|d(x,y) - d_{\mathrm{T}}(x,y)| > \delta) \cdot 100 \tag{17}$$

Both MAE and PBP metrics are utilised for evaluation, with values normalised for simplicity.

3.3. Elemental Image Compared to Viewpoint Image

VPI and EI are two image structures that can be obtained from Holoscopic images. Previous research has shown significant results in estimating disparity maps utilising VPIs. VPIs can be created by extracting a single pixel from each EI and arranging them in a tiled manner. However, the process of extracting VPIs does not consistently provide ideal images, unlike the VPIs found in synthetic datasets and those obtained from Lytro (The camera's performance was hindered by extensive pre-processing, resulting in slow performance.). The production of these images involves a significant amount of pre-processing. Occasionally, these procedures may require choosing a group of pixels instead of just one, employing shift and integration techniques, and utilising other methodologies to remove artefacts.

As depicted in Fig. 14, the Holoscopic images used have undergone calibration and rectification, ensuring that the grid of the EIs aligns perfectly to extract the VPI images accurately. VPI images are extracted using traditional methods, obtaining one pixel per EIs. Fig.15 displays three extracted VPIs from different locations. These images exhibit lower clarity, higher noise, and reduced resolution when compared to the images typically obtained from publicly available VPI datasets that have undergone extensive pre-processing. The difference in clarity between the EIs and VPIs can be seen in Fig. 14 and Fig. 15.

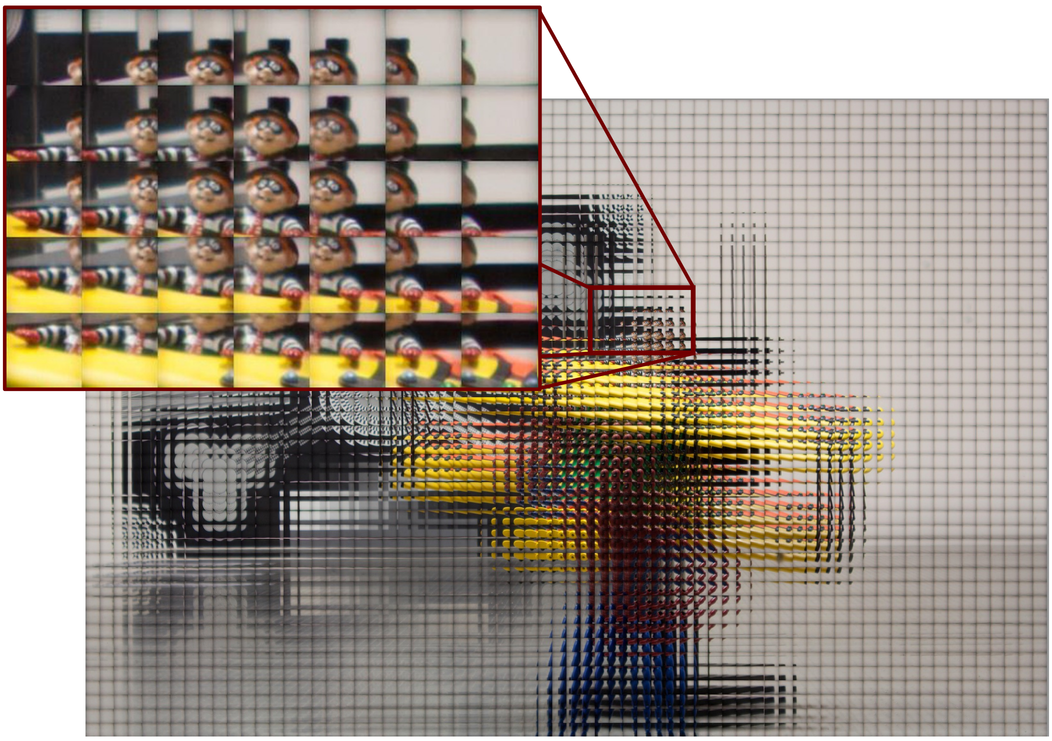


Figure 14. The Holoscopic image was calibrated and rectified, resulting in a total of 68 × 45 EIs, with each EI measuring 74 × 74 in size

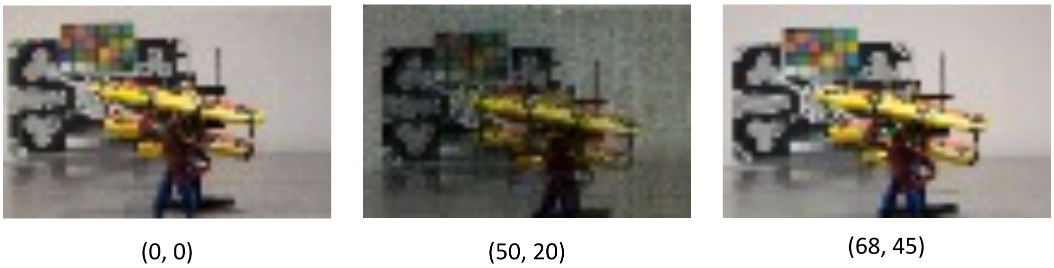


Figure 15. VPIs extracted from three different positions (0, 0), (50, 20), (68, 45)

Utilising pixel patches instead of single pixels to extract VPIs during pre-processing might lead to better outcomes, as demonstrated in Fig. 16. However, increasing the size of the extracted patch leads to a decrease in angular information, as the number of VPIs obtained is dramatically reduced. The number of VPIs is directly related to the resolution of the EI, which represents the amount of angular information captured. Moreover, while examining Fig. 16, it is apparent that the images require additional pre-processing to enhance the outcome. The process of obtaining VPIs also results in a substantial rise in the image generation time. This process can become particularly cumbersome when dealing with Holoscopic videos.

Figure 16. Holoscopic image, Spiderman: (64x34 MLA) 5160 × 2743 and sample images from different VPIs retrieved using patch sizes ranging from 5x5 pixels to 21x21 pixels (p). As seen in the extracted VPIs, they still exhibit some artefacts.

The disparity map was obtained from the EIs of the dataset captured by the Brunel Holoscopic camera using our approach, and subsequently obtained from the extracted VPIs. The disparity maps obtained from VPIs are then transformed to generate EIs, enabling a comparison between EIs with direct disparity estimates and EIs with disparity estimations derived from VPIs as seen in Fig. 17. The closeup crops from the Holoscopic image reveal that the disparity calculated by the EIs is distinct and clear, whereas the EIs obtained from the disparity generated by the VPIs are distorted and ambiguous As depicted in Fig. 18.

The disparity map, evaluated by the consistency check metric, utilises the entire raw Holoscopic image to optimise efficiency and minimise the amount of time and effort required. However, the evaluation of disparity using an edge-preserving approach is conducted between individual EIs. This metric is capable of detecting both the grid of EIs and the edges of the features within them. By utilising individual EIs, more reliable results can be obtained.

The edge alignment bar graph in Fig.19 (top) illustrates the MAE values for 12 distinct raw real Holoscopic images, which range between approximately 0.352 and 0.781. The changes seen can be attributed to disparities in scene, texture, colour, and complexity throughout the images. The results generally show lower values (better) in comparison to the edge-alignment metric results derived using VPIs disparity where the range of values

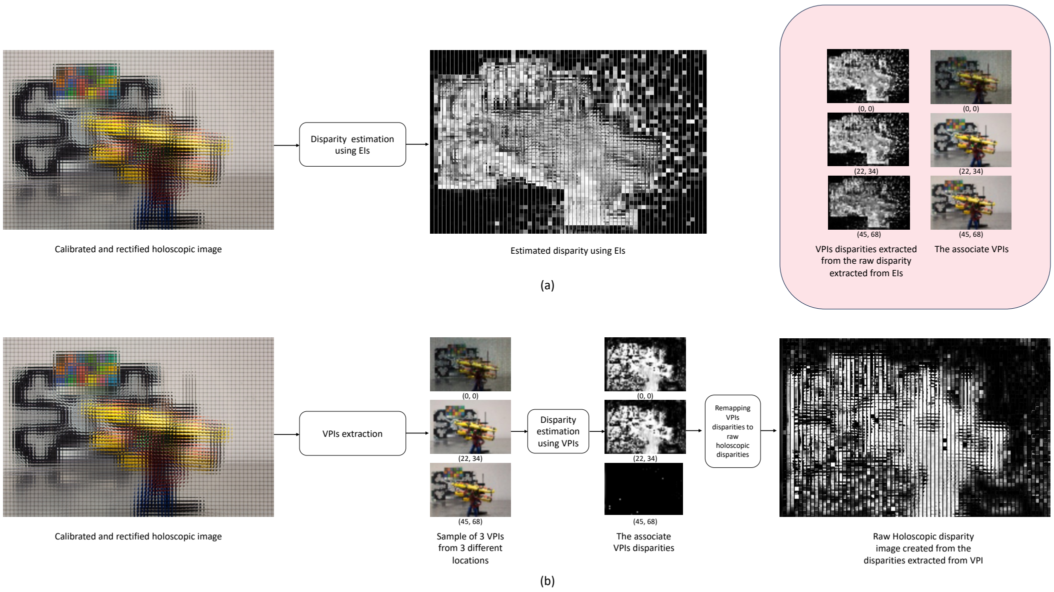


Figure 17. Disparity map derived from EIs and VPIs. (a) The disparity is calculated directly from the EIs using the raw Holoscopic image. Within the red-coloured box, there are a few extracted VPI disparities from the EI-based disparity. Their clarity is compromised by the low resolution. (b) VPIs are extracted from calibrated and rectified Holoscopic images, and the disparity map is obtained from them. These VPI disparities are then mapped back to EIs, allowing for a comparison between VPI-based and EI-based disparity maps.

for different images is approximately 0.498 to 0.797. Overall, EIs demonstrate better results in comparison to VPIs, with an average MAE of approximately 0.523, whereas VPIs have an average MAE of approximately 0.681, which can be viewed in the averaged bar "All".

The bar graph depicted in Fig.19 (bottom) illustrates the range of values for the consistency check metric derived from the disparity of EIs and VPIs. The values vary between approximately 0.334 and 0.756. The results exhibit lower values when compared to the consistency check metric results generated using VPIs disparity where the values range between approximately 0.548 and 0.881 for different images. Overall, EIs yield better results compared to VPIs, exhibiting average values of roughly 0.520, while VPIs demonstrate an average value of around 0.731. A selection of 4 raw Holoscopic images is shown in Fig. 20. Simple scenes were captured to compare the disparity of EIs vs VPIs directly rather than assessing the algorithm in a complicated scene configuration.

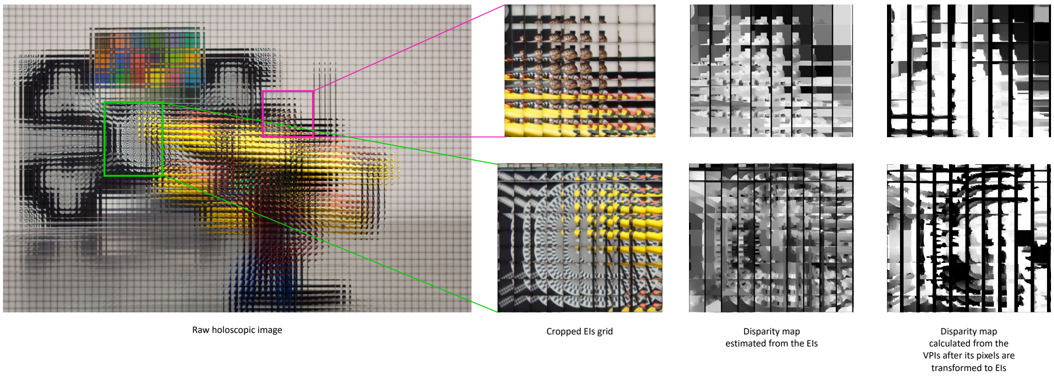


Figure 18. This is a close-up view of a raw Holoscopic image, along with the disparity maps derived from EIs and VPIs. The disparity map created from EIs has greater clarity compared to the one derived from VPIs.

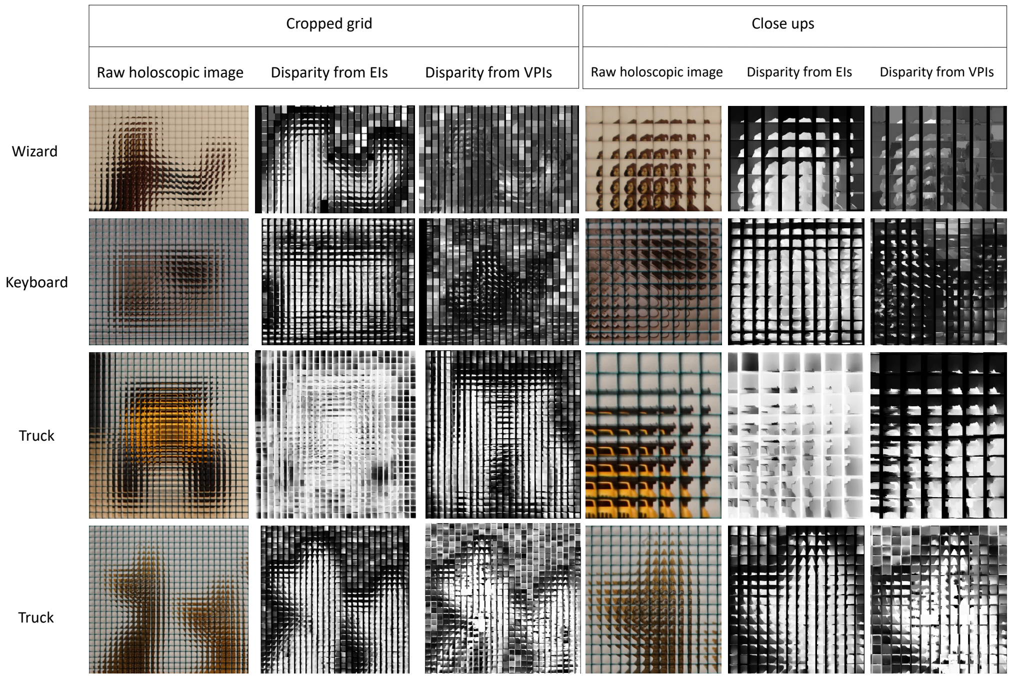


Figure 20. The algorithm was tested on a total of 12 real Holoscopic images. This is a collection of four images showcasing close-up sections to illustrate the disparity between EIs and VPIs.

3.4. Elemental Image Compared to Viewpoint Image Resolution

The algorithm's performance was assessed by utilising 24 synthetic Holoscopic images with five distinct EI resolutions: 20×20, 40×40, 60×60, 80×80, and 100×100 as shown in Fig. 21. MAE and PBP were calculated for all resolutions. As depicted in Fig.22, an increase in EI's resolution does not consistently result in improved accuracy. EIs with a high resolution are expected to lead to a high score. Yet, the clarity of the EIs relies on the clarity of the produced VPIs. Smaller EIs typically originate from VPIs with higher resolutions compared to those larger EIs (trade-off in resolution), which allows for more information to be presented in the EIs. This ultimately leads to a sharper image, as seen in Table 1. This table displays a single EI from various scales. Although the EI with a resolution of 100x100 is larger, it is noticeable that the circles on the dice in the EI with a resolution 60×60 are more defined and sharper. As the scale increases, the EI loses more information, resulting in the presence of noisy features. Future research can employ this

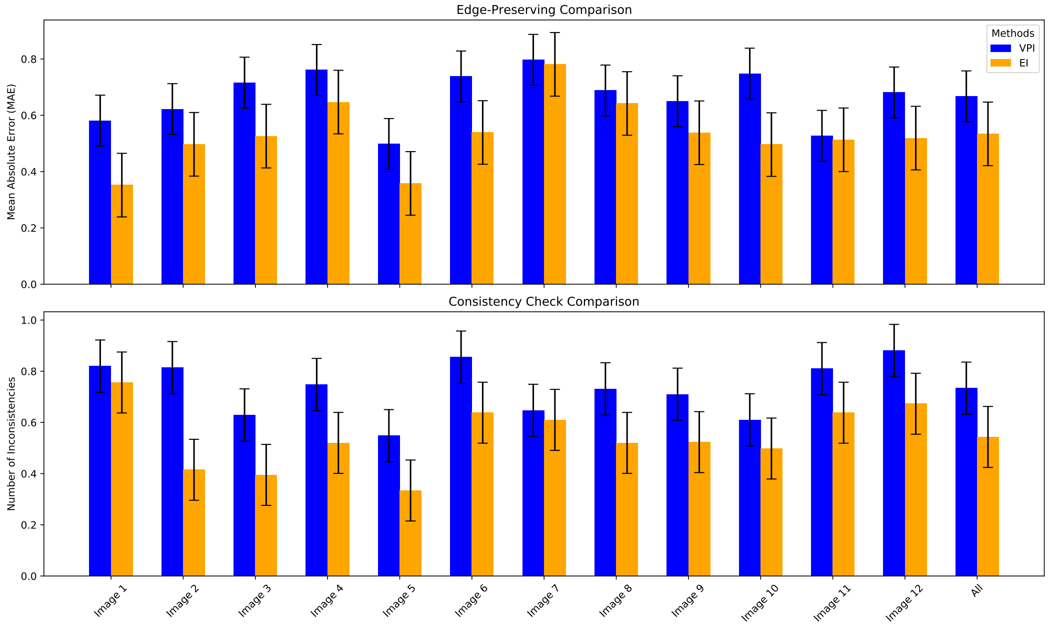


Figure 19. The bar graphs display the edge-alignment matrices (top) and consistency check matrices (bottom) calculated from 12 raw Holoscopic images captured by the Brunel Holoscopic camera. The averaged result is labelled as "All". EIs generally outperform VPIs, as seen by their lower average MAE and consistency check metric.

dataset with many resolutions to construct the multi-resolution pyramid, thereby capturing all the accessible information at each level of resolution.

EI Slice Resolution	20x20	40x40	60x60	80x80	100x100
Original Resolution					
Scaled-Down (20x20)					
Scaled-Up (100x100)					

Table 1. EIs of three different scales, original, down-sampled, and up-sampled.

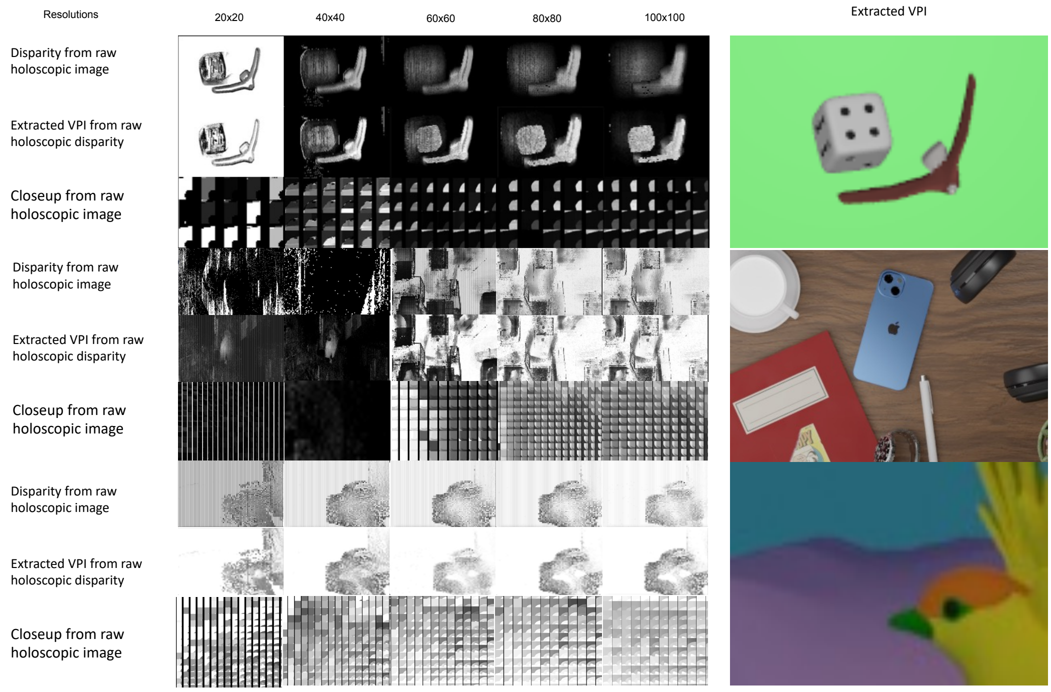


Figure 21. Example of three Holoscopic images alongside their calculated disparities at various resolutions. Observing the disparity from the low-resolution images is difficult. Consequently, close-up views are offered.

As seen in Fig. 22, the MAE values for the methodology across different resolutions reveal varying degrees of accuracy. The MAE for the 20×20 resolution ranges between 0.673 and 0.804, suggesting a significant amount of errors. For the 40×40 resolution, the MAE ranges between 0.613 and 0.755, indicating significantly enhanced performance in comparison to the 20×20 resolution. The 60×60 resolution's MAE ranges from 0.430 to 0.650, demonstrating a significant improvement in accuracy compared to the lesser resolutions. The MAE for the 80×80 resolution varies between 0.419 and 0.625, indicating a higher level of precision. At a resolution of 100×100, the MAE varies between 0.462 and 0.640, suggesting a somewhat lower level of precision compared to the 80×80 resolution.

The PBP for the 20×20 resolution ranges from 68.8% to 86.1%, suggesting a significant presence of bad pixels. The PBP of the 40×40 resolution falls within the range of 72.7% to 87.0%, indicating comparable performance to that of the 20×20 resolution. Moving to 60×60 resolution, the range is from 49.4% to 66.9%, suggesting a significant reduction in bad pixels compared to the lower levels. With an 80×80 resolution, the PBP falls between 39.6% and 64.1%, indicating a significant enhancement in performance and a reduction in the number of bad pixels. Finally, at 100×100 resolution, the range is from 44.7% to 64.3%, exhibiting accuracy that is slightly lower than 80×80 in accuracy.

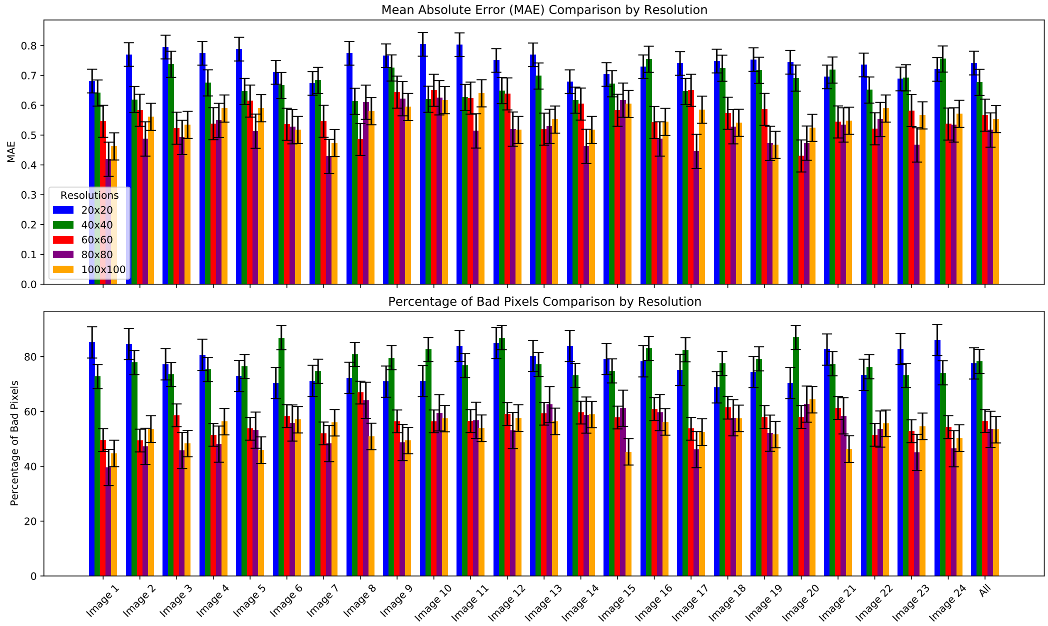


Figure 22. The bar graph illustrates the performance of disparity calculating at 5 different resolutions. The graph shows that the EIs achieve the highest level of precision at a resolution of 80×80, followed by 100×100.

408 Images with low resolution, such as 20x20 and 40x40, still have noticeably reduced accuracy. This is because achieving accurate disparity typically requires a combination of a wide baseline and a high-resolution image. Since larger EIs demonstrate a greater baseline and a reduced number of texture-less EIs, as depicted in Fig. 23 the result in high-resolution EIs re better than the accuracy in low-resolution images.

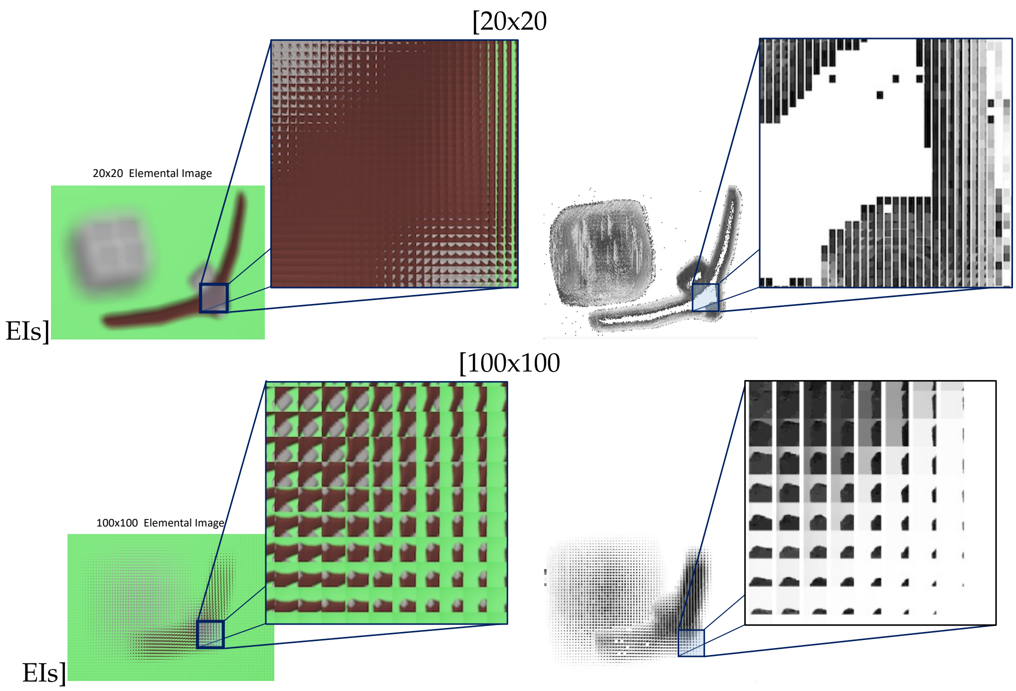


Figure 23. (a) shows the output of 20x20 pixel EIs with a significant texture-less area, leading to an incorrect disparity computation. In (b), the outcome of 100x100 EIs taken from the same point with a wider baseline and a larger portion of the objects presented leads to a more accurate disparity

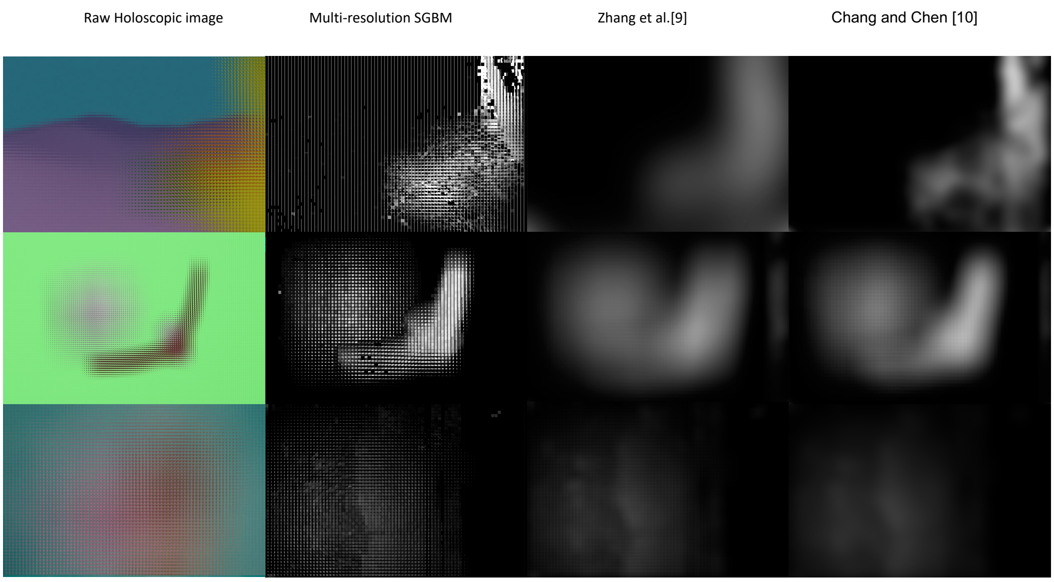


Figure 24. The result from bother deep learning algorithms result in blurry and undefined EIs compared to our result

3.5. Comparative Analysis of Stereo-Matching Networks

The results from raw Holoscopic images (EIs) were also compared against two state-of-the-art deep learning stereo matching algorithms: Zhang et al. [23] and Chang and Chen [24]. Zhang et al. [23] proposed a technique to enhance the generalisation abilities of stereo-matching networks. Their main objective was to maintain the consistency of features between corresponding pixels. Their methodology combines pixel-level contrastive learning with a stereo-selective whitening loss to enhance the consistency of features across various domains. This technique is highly versatile and may be easily integrated into pre-existing networks without any disruptions.

Chang and Chen [24] employ supervised learning and convolutional neural networks (CNNs) to address the task of estimating disparities from stereo image pairs. They proposed a Pyramid Stereo Matching Network (PSMNet) as an alternative to the patch-based Siamese networks commonly employed in current architectures. The PSMNet overcomes the limitation of incorporating contextual information in uncertain regions by incorporating spatial pyramid pooling and a 3D CNN.

Both of the pre-trained models were used to extract disparity from all 24 raw Holoscopic images in the dataset choosing 80x80 resolution based on the accuracy level from the previous section. These results were then compared with those obtained from this paper's method applied to the same dataset. The disparity outcome of a basic EI of both deep-learning models resulted in highly blurred and undefined results as seen in Fig. 24. These outcomes can be attributed to various factors including the dissimilar characteristics of the higher-resolution stereo images utilised for training the models developed by Zhang et al. [23] and Chang and Chen [24] compared to the low-resolution and low-texture EIs. Therefore, when these models are employed on the EIs, they struggle with accurately capturing intricate details. Furthermore, the efficacy of these models is greatly influenced by their specific architecture, particularly Chang and Chen [24]'s PSMNet, which further reduces the resolution of low-resolution EIs, resulting in unsatisfactory outcomes.

As depicted in Fig. 25, MAE values for this paper's method ranged from 0.419 to 0.625, which were considerably lower than the MAE values reported by Zhang et al. [23] ranging from 0.637 to 0.801 and Chang and Chen [24] ranging from 0.686 to 0.798. Regarding the PBP, this chapter's method achieved percentages ranging from 39.6% to 64.1%, which indicates superior performance. In comparison, Zhang et al. [23] and Chang and Chen

[24] obtained greater percentages, with ranges of 63.3% to 78.7% and 65.7% to 77.9%, respectively.

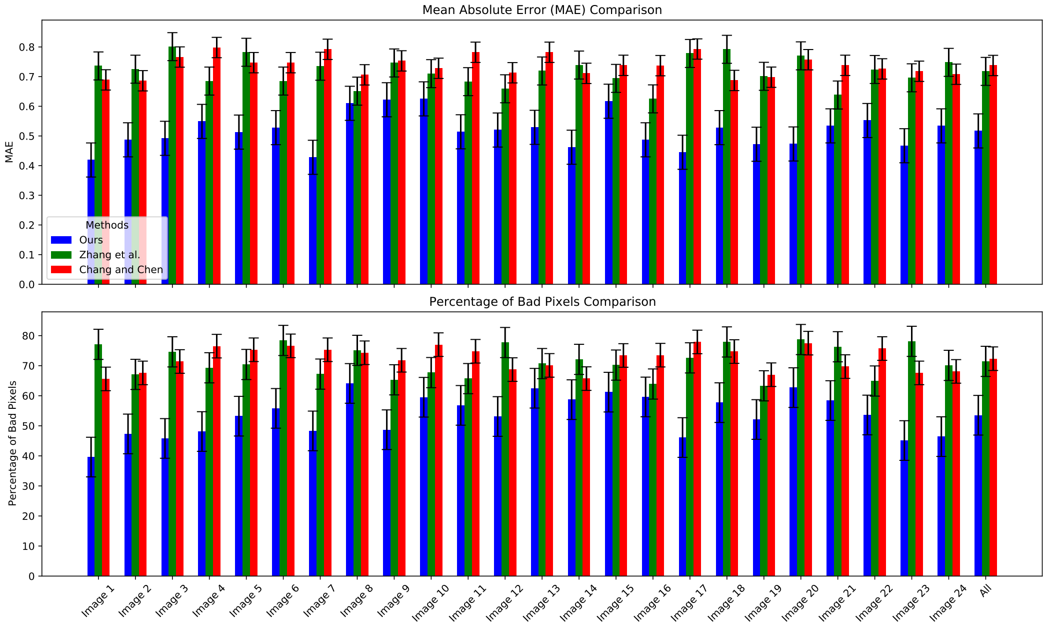


Figure 25. The bar graph depicts the comparative performance of the extracted disparity to Zhang et al. [23] and Chang and Chen [24], demonstrating that our method outperforms both methods algorithms

4. Conclusions

The critical component of Holoscopic technology is MLA, which mimics the effect of multiple cameras while requiring only a single lens and a single sensor design. For this reason, the micro-image, or EI, formed behind each micro-lens retains unique details about the scene's lighting, including its direction, colour, and intensity. Contrary to conventional methods for estimating disparity maps by employing extracted and up-sampled viewpoint images (VPIs) with spatial information, a novel method for generating a disparity map using EIs angular information is introduced. The utilisation of EIs for disparity estimates, as opposed to VPIs, has not undergone thorough investigation. The objective of this study is to investigate the practicality of using angular perspective data instead of spatial orthographic data to estimate disparity. Using VPIs requires extracting images, which can be a time-consuming task. Moreover, there is a potential for distortion to arise post-extraction due to lens aberrations, the convergence of light rays, and out-of-focus objects.

To compute disparity from the EIs, a semi-global block-based matching technique is utilised due to its flexibility. A pre-processing phase is conducted to improve the quality of EIs, which frequently exhibit low resolution and a lack of texture. The approach consists of two primary stages: noise reduction by bilateral filtering and contrast enhancement via histogram equalisation. The disparity between EI pixels is computed utilising the Semi-Global Block Matching (SGBM) technique, which is enhanced by implementing a multi-resolution approach to overcome the limitations in EI resolution. This procedure involves creating a multi-scale pyramid of EIs to accurately capture intricate details at different scales, while also utilising a content-aware analysis to adaptively adjust the SGBM window size configurations. This ensures a thorough evaluation of variations in texture and complexity at many levels within the EIs. Ultimately, a weighted least squares (WLS) filter is employed to further enhance the optimisation process. Furthermore, the presence of incorrect background EIs is identified and rectified by employing a background mask and adjacent EIs that include precise background data.

The investigation has determined that the proposed technique has effectively produced disparity maps that exceed the accuracy of VPIs in real images and outperform two advanced deep-learning algorithms. The approach was analysed using various EI resolutions to determine the optimal resolution.

Author Contributions: Conceptualization, R.S.; methodology, B.A; software, B.A.; validation, B.A., H.M., and R.S.; formal analysis, B.A.; investigation, B.A.; resources, B.A.; data curation, B.A.; writing—original draft preparation, B.A.; writing—review and editing, B.A. and H.M.; visualization, B.A.; supervision, H.M.; project administration, B.A.. All authors have read and agreed to the published version of the manuscript.

Funding: Not applicable.

Data Availability Statement: Data Availability Statement: The data presented in this study are available on request from the 350 corresponding author. The data are not publicly available due to commercial privacy.

Conflicts of Interest: The authors declare no conflicts of interest

References

1. Fatah, O.A.; Lanigan, P.M.; Aggoun, A.; Swash, M.R.; Alazawi, E.; Li, B.; Fernandez, J.; Chen, D.; Tsekleves, E. Three-dimensional integral image reconstruction based on viewpoint interpolation. In Proceedings of the 2013 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB). IEEE, 2013, pp. 1–4.
2. Lumsdaine, A.; Georgiev, T.; et al. Full resolution lightfield rendering. *Indiana University and Adobe Systems, Tech. Rep* **2008**, *91*, 92.
3. Lumsdaine, A.; Georgiev, T. The focused plenoptic camera. In Proceedings of the 2009 IEEE International Conference on Computational Photography (ICCP). IEEE, 2009, pp. 1–8.
4. Georgiev, T.; Lumsdaine, A. Reducing plenoptic camera artifacts. In Proceedings of the Computer Graphics Forum. Wiley Online Library, 2010, Vol. 29, pp. 1955–1968.
5. Almatrouk, B.; Meng, H.; Swash, R. Elemental Images Labelling and Grouping to Minimise Disparity Error in Texture-less Regions of Holoscopic Images. In Proceedings of the 2023 8th International Conference on Image, Vision and Computing (ICIVC). IEEE, 2023, pp. 531–536.
6. Ng, R.; Levoy, M.; Brédif, M.; Duval, G.; Horowitz, M.; Hanrahan, P.; et al. Light field photography with a hand-held plenoptic camera. *Computer Science Technical Report CSTR* **2005**, *2*, 1–11.
7. Kinoshita, T.; Ono, S. Depth estimation from 4D light field videos. In Proceedings of the International Workshop on Advanced Imaging Technology (IWAIT) 2021. SPIE, 2021, Vol. 11766, pp. 56–61.
8. Mousavi, M.; Khanal, A.; Estrada, R. Ai playground: Unreal engine-based data ablation tool for deep learning. In Proceedings of the International Symposium on Visual Computing. Springer, 2020, pp. 518–532.
9. Tankus, A.; Kiryati, N. Photometric stereo under perspective projection. In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1. IEEE, 2005, Vol. 1, pp. 611–616.
10. Park, J.H.; Baasantseren, G.; Kim, N.; Park, G.; Kang, J.M.; Lee, B. View image generation in perspective and orthographic projection geometry based on integral imaging. *Optics Express* **2008**, *16*, 8800–8813.
11. Eagle, R.; Hogervorst, M. The role of perspective information in the recovery of 3D structure-from-motion. *Vision research* **1999**, *39*, 1713–1722.
12. Thomas, G.A.; Stevens, R.F. Processing of images for 3D display, 2004. US Patent 6,798,409.
13. Hirschmuller, H. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence* **2007**, *30*, 328–341.
14. Elad, M. On the origin of the bilateral filter and ways to improve it. *IEEE Transactions on image processing* **2002**, *11*, 1141–1151.
15. Buades, A.; Facciolo, G. Reliable multiscale and multiwindow stereo matching. *SIAM Journal on Imaging Sciences* **2015**, *8*, 888–915.
16. Miangoleh, S.M.H.; Dille, S.; Mai, L.; Paris, S.; Aksoy, Y. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9685–9694.
17. Keys, R. Cubic convolution interpolation for digital image processing. *IEEE transactions on acoustics, speech, and signal processing* **1981**, *29*, 1153–1160.
18. Almatrouk, B.; Meng, H.; Swash, M.R. Disparity estimation from holoscopic elemental images. In Proceedings of the The International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery. Springer, 2020, pp. 1106–1113.
19. Liu, W.; Chen, X.; Shen, C.; Liu, Z.; Yang, J. Semi-global weighted least squares in image filtering. In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5861–5869.
20. Alzu'bi, A.; Amira, A.; Ramzan, N. Semantic content-based image retrieval: A comprehensive study. *Journal of Visual Communication and Image Representation* **2015**, *32*, 20–54.

21. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G.R. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the ICCV. Citeseer, 2011, Vol. 11, p. 2.
22. Almatrouk, B.; Meng, H.; Aondoakaa, A.; Swash, R. A New Raw Holoscopic Image Simulator and Data Generation. In Proceedings of the 2023 8th International Conference on Image, Vision and Computing (ICIVC). IEEE, 2023, pp. 489–494.
23. Zhang, J.; Wang, X.; Bai, X.; Wang, C.; Huang, L.; Chen, Y.; Gu, L.; Zhou, J.; Harada, T.; Hancock, E.R. Revisiting domain generalized stereo matching networks from a feature consistency perspective. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13001–13011.
24. Chang, J.R.; Chen, Y.S. Pyramid stereo matching network. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5410–5418.