An Occlusion Aware Five-view Stereo System and Its Application in Video Post-production

Changan Zhu[a], Taryn Laurendeau and Chris Joslin[b]
School of Information Technology, Carleton University, 1125 Colonel By Drive, Ottawa, Canada

Keywords: Stereo Vision, Disparity Enhancement, Image and Video Processing, Post Production, Visual Effects.

Abstract: Extracting the live-action elements from videos has been a time-consuming process in the post-production pipeline. The disparity map, however, shows the order of elements in a scene by indicating the distance between the elements and the camera, which could potentially become an effective tool for separating videos into ordered layers and preserving the 3D structure of the elements. In this research, we explored the possibility of simplifying the live-action video element extraction technique with disparity sequences. We developed a five-view disparity estimation and enhancement system with a two-axis setup that helps reduce the occlusions in stereo vision. The system is independent from temporal reconstruction hence is compatible with both dynamic and stationary camera paths. Our results show that the disparities from our system have visually and quantitatively better performance than the traditional binocular stereo method, and its element extraction result is comparable with the existing mature matting techniques in most cases. Ideally, the system design could be applied in cinematography by replacing the center camera with a cinematographic camera, and the output can be used for video object extraction, visual effects composition, video's 2D to 3D conversion, or producing the training data for neural-network-based depth estimation research.

1 INTRODUCTION

Video element extraction is one of the most widely used techniques in the current post-production pipeline. Media such as movies, television series, and advertisements all use the technique to either replace the background or composite visual effects (VFX) into the original scene.

Chromakey, also known as color keying, is an effortless approach for element extraction that the industry has loved for decades. However, the element extraction result from it can be easily influenced by lighting conditions and the foreground objects; it also requires a large scale of preliminary setup and is restricted by the environment (Gvili et al., 2003). For areas that are not covered by green screen or situations that the background needs to be preserved, Rotoscoping would be used. Yet the extraction quality of Rotoscoping is highly dependent on the manual efforts, it has a simple mechanism and is widely used but also high-cost in terms of time consumption and workload (Li et al., 2016). Other element extraction approaches in the industry, such as motion tracking and auto-roto-painting, are either constrained by camera movement or objects' texture similarity. Additionally, the mattes acquired from these techniques cut out the elements as 2D planes, which only slice a scene into few layers and provides limited composition capability. On the other hand, the disparity map reflects an element's position in a 3D scene, presenting the element's 3D structure and its order in the environment. The post-production pipeline usually adapts depth maps from CG objects to enhance their composition results in live-action scenes (FXGuide, 2014).

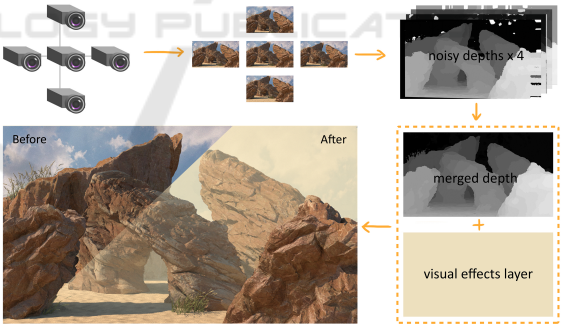

Figure 1: System procedure.

[a] https://orcid.org/0000-0002-2980-813X
[b] https://orcid.org/0000-0002-6728-2722

Zhu, C., Laurendeau, T. and Joslin, C.
An Occlusion Aware Five-view Stereo System and Its Application in Video Post-production.
DOI: 10.5220/0010884400003124
In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 5: VISAPP*, pages 844-854
ISBN: 978-989-758-555-5; ISSN: 2184-4321

Hence, having the disparity information of the live-action objects might also effectively assist the matting and visual effects composition. According to our literature review, existing accurate disparity estimation methods such as LiDAR, Time-of-Flight have limits in the distance and usually provide sparse depth data that requires completions (Ma et al., 2019); Multi-view methods like Structure-from-Motion (SfM) are normally based on a static world assumption (Vijayanarasimhan et al., 2017), thus are hardly applicable on dynamic objects. Supervised and semi-supervised monocular depth estimation networks require ground truth depth input that involves intensive labor work; while the unsupervised networks rely on adjacent frames for depth refinement that a moving camera shot is often needed (Ming et al., 2021). Stereo vision, as one of the most well-studied depth estimation techniques, is more flexible with the challenges from spatial density, non-stationary elements, and camera paths while its accuracy and visual output suffer from matching issues brought by feature mismatch and occlusion (Bleyer and Breiteneder, 2013). Therefore, we proposed a stereo system (see in Figure 1) that aims to simplify the element extraction in post-production and solve the occlusion problem in stereo matching. Our system acquires four stereo pairs from horizontal and vertical axes by perceiving a scene from five views, then centripetally align and merge the disparity results. In this case, we can compensate the occlusions from one axis with the information from another axis. If we compare the occluded area to shadow, our system reduces the occluded areas like a shadow-less lamp. We evaluated our system quantitatively by comparing its results with traditional binocular stereo. Also, we provided a visual comparison of the matting results from our method and industry standard method. Accordingly, we contributed:

- A depth acquisition system that is fully compatible with two non-overlapped baseline axes, which generates quantitatively and visually better disparity than the state-of-the-art binocular stereo methods.
- A two-axis stereo rectification solution.
- A stereo dataset that provides stereo pairs from five views and two axes with corresponding ground truth.
- A matting method that could simplify the element extraction process and provide a 3D matting option for more realistic visual effects.

2 RELATED WORKS

2.1 Element Extraction Approaches

In our research, the term element extraction refers to the process of pulling target elements out from a video, which relates to concepts such as video matting and segmentation. Available matting solutions can be categorized into color keying, alpha matting, and other innovative methods such as depth and defocus.

2.1.1 Color Keying

Color keying provides high-quality matting results for evenly lit-up objects that are filmed against a blue or green screen. Commercial tools such as the Ultimatte (Blackmagic, 2021) and Keylight (Foundry, 2021) both provide outstanding foreground isolation results. However, the method is restricted by the environment, the lighting condition, and foreground colors.

2.1.2 Alpha Matting

Alpha matting is well discussed in academia but has not been widely applied in the industry as it provides limited robustness when dealing with fuzzy edges and homogeneous regions. In pixel-sampling-based alpha matting, it always requires well-specified trimaps (a map showing the definite foreground, definite background, and the blended regions with both foreground and background pixels) for accurate matting results, and the result can be noisy when the provided trimap is rough or when the input image contains highly textured regions (Chuang et al., 2002). The pixel-affinity-based alpha matting method has better performance in terms of matte continuity and complex background while it usually requires manual annotation to help reducing errors from propagation (Levin et al., 2008).

2.1.3 Depth Keying

The idea of using depth as an element extraction tool was present early. Kanade et al. (Kanade et al., 1996) addressed the possibility of using depth to isolate elements from a video clip automatically. Givili et al. (Gvili et al., 2003) suggested the concept of "depth keying" that extracts foreground objects from the background using depth. After this, approaches that use depth as guide information for more accurate matting results were presented (Wang et al., 2012; Lu and Li, 2012).

2.1.4 Other Approaches

Researchers also propose many other innovative matting approaches such as segmentation, contour tracking, and light field. Li et al. (Li et al., 2005) proposed to extract video objects with a 3D graph-cut-based segmentation and a tracking-based local refinement. Chung and Chen (Chung and Chen, 2009) present a video segmentation method with Markov Random Field (MRF) based contour tracking. The limitations of these methods are that most of them require manual annotations, and discontinuities can be introduced due to the feature points lying beyond the image boundaries. Wu et al. (Wu et al., 2017) mentioned in their work about the matting problem that could be solved with light field, but also pointed out the restrictions from computational resources and running time for light-field data.

2.2 Depth Estimation and Enhancement

Traditional vision-based stereo has advantages in fewer environment constraints, consistent spatial resolution, high portability, and low cost, hence commonly used in the applications and research of depth estimation. However, since it relies on finding correspondence points from a stereo pair, errors are frequently arisen by correspondence mismatch and occlusion. To improve the accuracy of depth estimation, researchers have explored both hardware and software sides to produce or enhance disparities. For example, existing solutions such as using light and laser (Zhu et al., 2008; Silberman and Fergus, 2011; Ferstl et al., 2013), optical features (Jeon et al., 2015; Zhou et al., 2009; Tao et al., 2015), camera arrangement (Kanade et al., 1996; Honegger et al., 2017), different kinds of filters (Park et al., 2011; Yang et al., 2007; Barron et al., 2015), and most recent research with the assistance from neural networks (Park et al., 2018; Kopf et al., 2020).

2.2.1 Light & Laser

Methods such as Time-of-Flight and LiDAR are accurate within a certain distance and are used to create ground truth disparity for many depth-estimation related researches (Scharstein et al., 2014). However, both techniques create sparse data that requires completion (Ma et al., 2019). The distance limit and their synthesizing with the cinematographic camera (Nair et al., 2013) are also challenges lying ahead.

2.2.2 Optical Features & Camera Arrangement

Methods such as light field or camera arrays also provide exciting accuracy. To acquire depth from light-field, techniques such as pixel variations (Manakov et al., 2013; Heber et al., 2013), and graph-cut (Jeon et al., 2015) are commonly used, while since it does not model the occlusion boundaries, the performance could be limited in object boundaries (Ihrke et al., 2016) and noisy backgrounds, when it comes to handling dynamic objects, light-field-based methods could become very high-cost. Research using camera array were also frequently proposed in the 2000s and 2010s (Zhang and Chen, 2004; Fehrman and McGough, 2014; Tao et al., 2018). The method is different from light-field in the sampling patterns (Ihrke et al., 2016), while the techniques of acquiring depth with camera array are rather similar to light field (e.g., Depth-from-Defocus). Typically, camera arrays require large amount of cameras hence research have not investigated its application in a cinematographic context.

2.2.3 Neural Networks

Monocular depth estimation using neural networks does not require additional equipment, the results are also temporally consistent that could be applied in many daily life scenarios. However, the supervised and semi-supervised networks' training need ground truth depth input, which are costly to acquire; the unsupervised networks either require moving camera paths to acquire and refine depth basing on the adjacent frames, or need stereo matching results for the training (Ming et al., 2021).

2.2.4 Multi-view & Multi-baseline

Among all the depth enhancement technologies, the improvement from the hardware side seems to be unavoidable, especially when precision and quality are the primary goals. The multi-baseline stereo uses more than one set of camera pairs with different baselines to reduce occlusion and produce better depth (Honegger et al., 2017). Multi-view stereo, also known as structure from motion (SfM) (Schonberger and Frahm, 2016), acquires objects' 3D structure from a camera view's temporally change. Most multi-baseline and multi-view stereo methods could significantly improve the depth quality by adding cameras or number of frames (Schonberger and Frahm, 2016; Honegger et al., 2017). However, since the multi-view stereo relies on temporal scanning, the dynamic objects in a scene could corrupt the consistency thus need to be removed or masked (Klodt and Vedaldi,

2018; Luo et al., 2020). In the multi-baseline side, existing methods mostly arrange cameras on the same axis due to the constraints from epipolar geometry, where it requires many cameras and various baselines to reduce the occlusions, while the direction of the occluded areas remain unchanged as the cameras are setup on only one-axis. Therefore, the capability of such system to improve depth quality is also limited. Better results might be achieved with fewer cameras and less variant in the baseline distances, but from two different axes. In which case, the occlusion that normally exists in one axis due to the single-direction camera arrangement could be compensated by the depth information acquired from the other direction.

3 METHODOLOGY

Since the target scenario of our system is video post-production, the disparity generated from our system should have well-defined depth information at the target's edges for a relatively good matting result, and a smooth depth gradient that could accurately reflect the target's 3D structure for the purpose of visual effects composition. Additionally, we expect the system to automatically generate disparity sequences without excessive manual interventions.

3.1 System Overview

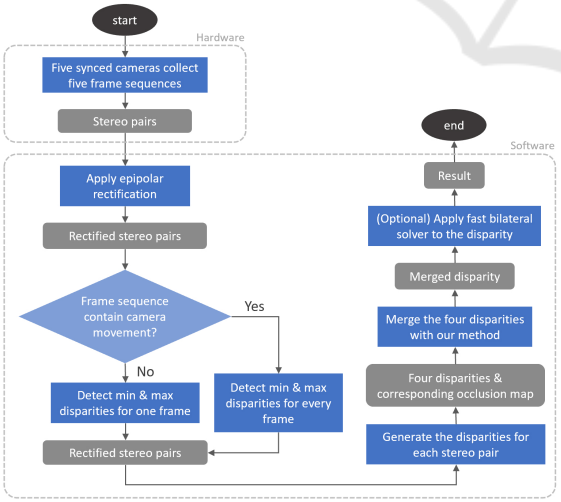

Figure 2: System overview.

As shown in Figure 2, our system consists of a hardware setup from five computer vision cameras and a framework for the disparity generation and refinement. As each secondary camera captures a stereo pair with the primary camera, the system first conducts a two-axis rectification on each stereo pair, then estimate and merge the disparities to output an enhanced disparity map. Optionally, a bilateral filter could be applied to create smoother visual result.

3.1.1 Hardware

For the hardware setup (Figure 3), we used five computer vision cameras that are connected through GPIO cables with synchronized imaging parameters, trigger signal, and framerate. To deploy the five cameras with our two-axis design pattern, we modelled and 3D printed a mount that can fix the cameras; and a lens holder that could reduce unexpected rotation among cameras to prevent massive distortion from rectification.

Figure 3: Hardware setup.

3.1.2 Development

On the software side, we developed a framework containing a two-axis image rectification method, an implementation of cost-volume filter optimized Winner Takes All (WTA) disparity estimation (Hosni et al., 2012), and a disparity merging algorithm with helper modules.

Two-axis Rectification. To reduce the computational cost in the correspondence matching, the stereo-vision-based methods normally involve stereo pair rectification as a crucial step for reducing the cameras' intrinsic distortion and extrinsic positional deviation. Most multi-baseline stereo methods only arrange cameras on one axis (Kang et al., 2008; Yang et al., 2014) due to the constrain of epipolar geometry. However, as our research set up the cameras on two axes, we looked into non-epipolar methods like correspondence-point-based rectification (Ota et al., 2009; Nozick, 2011). In the research of Nozick (Nozick, 2011), he considers the image rectification process as a rotation around the optical center and an update of the focal length, and simplified the problem to finding the relations between the original image and

the rectified image, which is represented as homography matrix H_i that satisfies the equation 1:

$$(H_i x_i^k)_y - y_k = 0 \quad (1)$$

where y_k is the average y-coordinate of the k^{th} rectified correspondence points, x_i^k is the correspondence points so the points are horizontally aligned to the same y-coordinate (Nozick, 2011). However, it also indicates that this method cannot be applied into our system as we compute the depth for the cameras in vertical directions by considering the y-axis as their baselines, hence their y_k would be the x_k of the horizontal cameras. In another word, the stereo pairs from the two axes would be rectified based on different y_k, thus we cannot rectify the images from both directions at the same time. Otherwise, the rectified stereo pairs would produce center disparities from different image planes, where the objects do not align and the results cannot be merged.

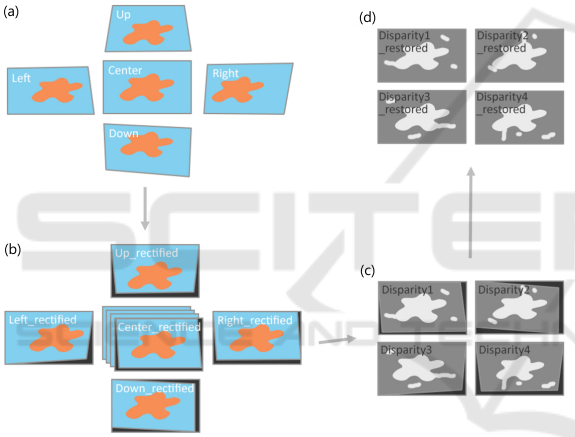

Figure 4: Our image rectification procedure.

Benefit from the pattern of our camera setup, we resolved the problem in a simple way. As shown in Figure 4, the two-axis rectification procedure follows: firstly (Figure 4 (a)(b)), using epipolar-geometrical-based method (Fusiello and Irsara, 2008) to rectify the images as four stereo pairs and keep a copy of the homographies that were applied on each pair; secondly (Figure 4 (c)), generate disparities from them as per pair, making sure the disparities always align to the rectified center image, now the four disparity maps are on different image planes; finally (Figure 4 (d)), calculate the reverse matrices for the previous homographies and apply them on each disparity, so the disparities are restored to the original center camera's image plane, where the object boundaries would align, and the disparities are ready to be merged.

Occlusion Aware Disparity Merging. To determine from which disparity map should we select the depth value and save to the result, we used the occlusion map as the guide information that regions occluded in one stereo pair are patched with the depth values from another pair, where the same region is not occluded hence the depth value is more reliable. We also introduced a threshold inspired by the ratio test in SIFT algorithm (Lowe, 2004) for filtering the outlier depth values hence produces disparities with less noises. The merging method is expected to fully use the disparity information from the four disparities, then compensate the disparity errors brought by occlusion in the merged result. The detail of the merging process is shown in Algorithm 1.

Algorithm 1: Occlusion Aware Disparity Merging.

Input: Sequence length *SeqLength*, threshold T, four disparity sequences $Disp1_i$, $Disp2_i$, $Disp3_i$, $Disp4_i$ and corresponding four occlusion sequences.
Output: The merged disparity sequence Mrg_i
$i = 0$;
repeat
  for *each point's depth value* $P_j, (j = 1,2,3,4)$ *in the four disparity maps* $Disp1_i$, $Disp2_i$, $Disp3_i$, $Disp4_i$ *and occlusion maps.* **do**
    if P_j *is not occluded in all four disparities or* P_j *is all occluded in the four disparities* **then**
      if *one of the* $P_j, (assume\, j = 1)$ *is* $(1+T)$ *times larger or* $(1-T)$ *times smaller than the other three pixels* **then**
        $Mrg_i = \text{mean}(P_j, (j = 2,3,4))$
      else
        $Mrg_i = \text{mean}(P_j, (j = 1,2,3,4))$
      end
    end
    if P_j *is occluded in three disparities, assume only the* P_1 *is not occluded* **then**
      $Mrg_i = P_1$
    end
    if P_j *is occluded in two disparities, assume* P_1, P_2 *are the two points not occluded* **then**
      $Mrg_i = \text{mean}(P_j, (j = 1,2))$
    end
    if P_j *is occluded in only one disparities, assume* P_4 *is the occluded point* **then**
      if *one of the* $P_j, (assume\, j = 3)$ *is* $(1+T)$ *times larger or* $(1-T)$ *times smaller than the other two points* **then**
        $Mrg_i = \text{mean}(P_j, (j = 1,2))$
      else
        $Mrg_i = \text{mean}(P_j, (j = 1,2,3))$
      end
    end
  end
  i ++;
until $i == SeqLength - 1$;

Helper Modules (Min and Max Distances Detector & Fast Bilateral Solver). In the disparity estimation procedure, it could speed up the correspondence matching process if with known min and max pixel distances among all corresponding points, as they constrain the searching distance of the algorithm. A moving camera shot or non-static element in the scene would require dynamic min and max disparities information for more efficient disparity estimation. To fit our system in such scenarios, we developed a simple min-max-disparities detector module based on SURF and FLANN matching, where we used SURF to compute the correspondence features from a stereo pair and FLANN to filter the outlier matches by using the Euclidean distances among the descriptors of features. To produce smoother visual output, we also adapted the fast bilateral solver from Barron and Poole (Barron and Poole, 2016) as an optional step in the framework, while we did not use the disparities processed by the solver in the quantitative evaluation to ensure the evaluation's accuracy.

3.2 Experiment

To compare the binocular method with our method, as well as to compare the matting results from the industrial standard toolbox and our disparity matting, we designed the experiment with two main comparisons:

1. A quantitative comparison (binocular disparity vs. our disparity, Chromakey matting vs our matting): using rendered image sequences from virtual environments to generate disparities and mattes, and evaluating their accuracy with rendered disparities and mattes ground truth;
2. A visual Comparison (binocular disparity vs. our disparity, Chromakey matting vs our matting): Using live-action image sequences, generating disparity and matte sequences, compare the disparity output with binocular stereo methods, and compare the matting result with industry matting tool;

For the binocular disparity, we selected the same disparity estimation method (Hosni et al., 2012) used in the system to ensure a consistent comparison.

3.2.1 Data Collection

To test and collect data with our five-camera setup, we designed both virtual and live-action environments for the data collection.

Computer Generated (CG) Data. The CG data are collected through Maya, where we can easily simulate various virtual scenes and deviations, generate ground truth depths and mattes for quantitative comparison. We created two virtual environments with the five-view camera setup (virtual cameras) as the system design. Each environment is split into three scenarios, which are the ideal, the semi-realistic, and the realistic scenarios. The three scenarios all have a stereo version for disparity accuracy comparison and a version with a green screen for the matting accuracy comparison (see in Table 1).

Table 1: The virtual environment setup for disparity and matting accuracy comparison.

	Chromakey version	Stereo version
Ideal	• No color-spill • No shadow on the screen • No objects with screen color	• Evenly lit up • Objects with obvious color difference • No camera rotation or translation
Semi-real	• No color-spill • Has shadow on the screen • Has objects with screen color	• Low light • Objects with similar color and texture • No camera rotation or translation
Realistic	• Has color-spill • Has shadow on the screen • Has objects with screen color	• Low light • Objects with similar color and texture • Has camera rotation and translation

Live-action Data. The live-action data are collected with our five-camera system. We recorded videos from eight scenes and selected the three best synchronized with no corrupted or skipped frames. Due to the fact that the ground truth for the live-action scenes cannot be generated with available experiment tools, they are shown in the visual comparison section and the accompanying video.

3.2.2 Metrics

In the experiment, we involved two metrics for evaluating our system output.

Disparity Accuracy. As applying the disparities from our system in post-production is our primary goal, the disparity accuracy is an essential factor that could reflect the system's performance. In the evaluation of our disparity accuracy, we followed the convention in existing disparity evaluation methods (Scharstein et al., 2014) that we chose Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) as two of our main measures. Since our objective is to apply the output to post-editing, where the quality is mainly evaluated through human visual system, we also included Structural Similarity Index (SSIM) as it effectively measures image dif-

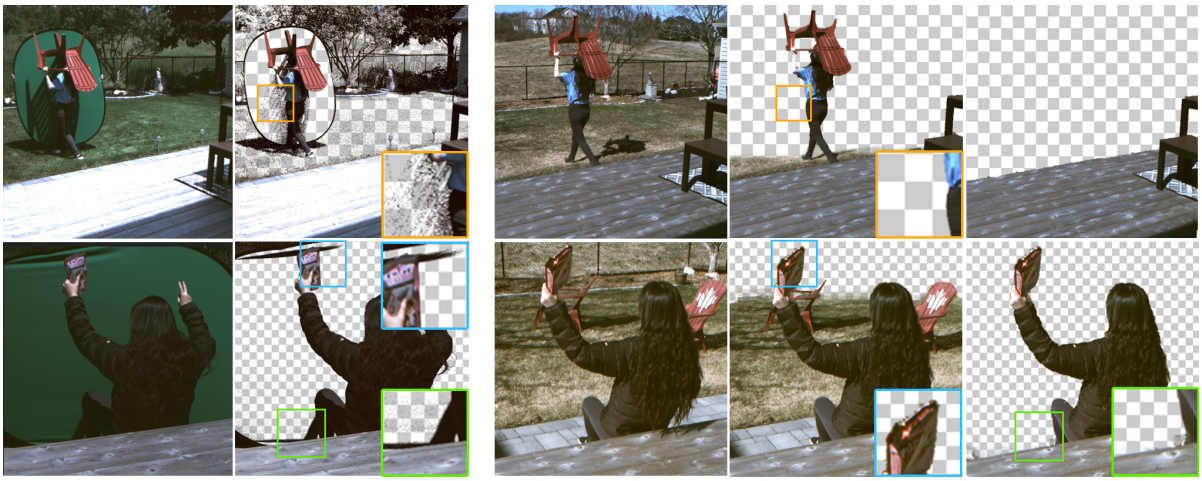

Figure 5: Visual comparison of matting between Chromakey and our result.

ference from a human visual perspective (Sara et al., 2019).

Matting Accuracy. While generating matte from our disparity bridges the research output and the post-production process, the matting accuracy is another key evaluating metric in our experiment. As is elaborated in Section 3.2, we compared the matte from our disparity with an industrial color keyer Keylight (Foundry, 2021). For the measures, we also used RMSE, MAPE, and SSIM.

We animated the virtual environments with movements applied on the objects or cameras for both disparities and matting accuracy evaluations. Every sequence for each scenario is 100 frames long. Our comparison took the mean value of each measure from the 100 frames instead of the values from only one frame.

4 RESULTS

4.1 Quantitative Comparison

4.1.1 Disparity Accuracy

As shown in Table 2, our method shows higher SSIM result in all scenarios and remains stable in the realistic scenario while the SSIM of the binocular method drops significantly when the camera relative rotation and translation are introduced.

The RMSE and MAPE are error metrics that show reverse behavior with SSIM. The disparity error gap between our method and the binocular method are relatively small in ideal and semi-realistic scenarios but widen when it comes to a realistic scenario.

Table 2: The disparity accuracy comparison result.

Measure	Method	Ideal	Semi-realistic	Realistic
RMSE ↓	Our method	**6.41**	**7.66**	**9.39**
	Binocular stereo	14.65	17.63	37.90
MAPE ↓	Our method	**0.91%**	**0.96%**	**1.30%**
	Binocular stereo	1.56%	1.78%	9.64%
SSIM ↑	Our method	**95.36%**	**94.88%**	**94.01%**
	Binocular stereo	93.12%	92.64%	78.50%

4.1.2 Matting Accuracy

The matting accuracy is also evaluated under ideal, semi-realistic, and realistic scenarios. Chromakey's matting results are separated into raw and adjusted versions for a more comprehensive comparison since Keylight might reach a high matting accuracy with some parameter changes. The raw version is simply to extract the green screen color from the video without any parameter adjustment, and the adjusted version is the best Keylight matting accuracy we can get by changing the parameters. The adjustment does not involve any garbage matte or post-editing.

As can be observed from Table 3, it is not surprising to see that adjusted Chromakey results show higher SSIM and lower error rates than ours in ideal and semi-realistic scenarios. However, our results provide better performance in the realistic scenario, and the unprocessed Chromakey matting results are always less accurate than ours, which means Chromakey has to involve manual adjustments for a decent output while our method performs well by running automatically.

Table 3: The matting accuracy comparison result.

Measure	Method	Ideal	Semi-realistic	Realistic
RMSE ↓	Our method	13.25	15.32	**16.43**
	Keylight raw	9.22	19.64	49.23
	Keylight adjusted	**7.88**	**11.43**	21.65
MAPE ↓	Our method	0.28%	0.37%	**0.44%**
	Keylight raw	0.76%	0.32%	14.45%
	Keylight adjusted	**0.16%**	**0.24%**	0.97%
SSIM ↑	Our method	98.80%	98.56%	**98.45%**
	Keylight raw	83.71%	76.57%	44.28%
	Keylight adjusted	**99.33%**	**98.71%**	96.05%

4.2 Visual Comparison

In the visual comparison for disparity, we generated depth sequences with both the binocular stereo method and our method. In the comparison (Figure 6), our output shows cleaner edges, smoother depth gradient, and has corrected many depth value errors brought by occlusion. We also provided a visual comparison of matting between Chromakey and our method.

Figure 6: Visual comparison of disparity between binocular stereo and our result.

For the depth-based-matting, we used our five-view system to capture the screen-free scenes; for the Chromakey-based matting, we replicated the movements from the screen-free scenes but filmed with green screen. The matting results (Figure 5) reveal that even on the green screen covered regions, it is hard to acquire a noise-free matte if there is a dense shadow on the screen or any object with screen color in the shots. Even the adjusted Chromakey matting results require a large amount of garbage matte due to green size limit and noises on the shadow area. On the contrary, our method only requires selecting a range of depth to create a matte. Also, we can select an arbitrary range of depth from the disparity sequence to mask out any part of the scene instead of only being able to extract objects with a green screen behind.

4.3 Composition Test

To further demonstrate the matting capability of our output, we composite visual effects into the videos and compared it with the composition result from Chromakey (more examples are included in the accompanying video).

As can be observed from Figure 7, since Chromakey only provides a two-layer matte that separates the foreground and background as two planes, we can only insert the effects as a single layer. However, with disparity sequence, the scenes could be sliced into multiple layers, where we can select a specific range to insert the effects. When compositing effects that the amount of its opacity is influenced by objects' depth (e.g., fire, explosion, sandstorm, fog), having the disparity information of certain element could help blend the visual effects into the live-action shot for more vivid composition result.

4.4 Discussion

According to the results from our experiment, our system generates disparity maps that have quantitatively and visually better performance than the traditional binocular disparity acquisition method, especially after the real-life challenges were added to the experiment. Comparing the matting results between our method and the commercial color-keying method, our disparity-based matting shows less accuracy in an edge-detail-preserving perspective, but provides outstanding matte accuracy within objects though adding with highly influential challenges such as camera's positional deviations and color inconsistency. Besides, our results reflect the depth gradient of objects, which provides the post-production pipeline with another option to composite effects regarding objects' 3D structure, and the process is fully automated.

5 CONCLUSION

This research presents a stereo system that perceives a scene from five views, where cameras from different

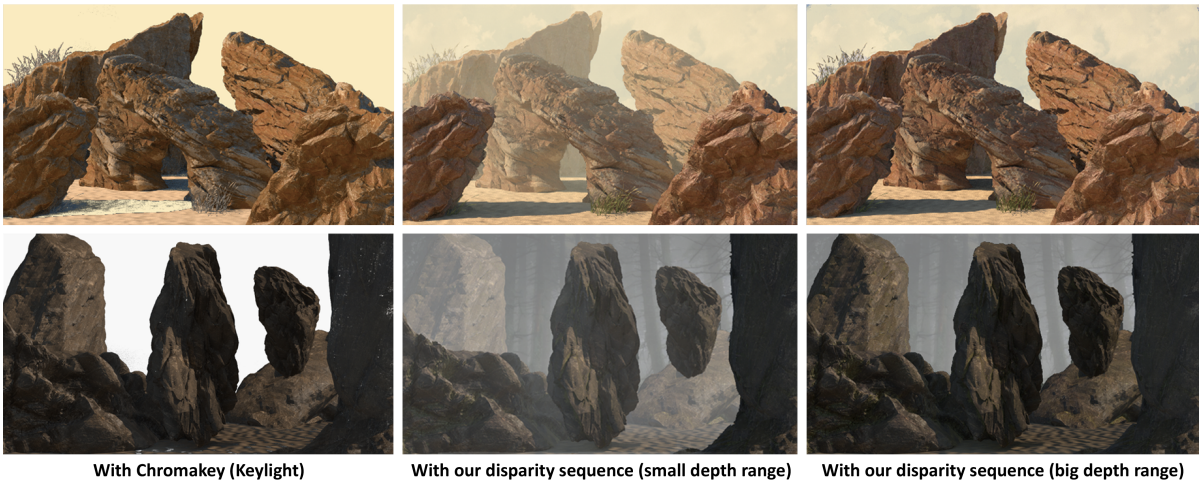

Figure 7: Insert effects into the virtual scenes.

directions fill the blind spots that other cameras cannot capture due to occlusion. The system is compatible with both stationary and dynamic camera shots and objects and has provided a new solution to live-action elements' extraction and visual effects composition. Throughout the research, we present a two-axis image rectification solution that efficiently reduced the difficulty of rectifying stereo pairs from various axes and have contributed the first stereo dataset that consists of five views in two axes, which are applicable in other multi-baseline or multi-view stereo researches.

Our research is not free of limitations. Since our research focuses more on solving the element extraction problem in the post-production, we compared the matting results from our method with the the industrial solutions. In the disparity side, we compared our method with binocular stereo, but did not compare with other multi-camera stereo methods, which should be added in future for a more comprehensive evaluation. We have not yet explored the disparity matting for transparent or semi-transparent objects; while it might be a frequently encountered task in real-life video filming scenarios, we shall explore the possibility of applying our method in such circumstances and replenish our system with new modules for more complicated situations.

In the future work, we would like to attempt synthesizing the stereo vision cameras and cinematographic camera, hence narrowing down the gap between the research and the industrial practices. On the software side, we would like to explore the integration of image segmentation for reducing matching errors caused by texture-less regions and the algorithm's lack of understanding of the semantics. Also, learning-based temporal smoothing could be introduced for more consistent depth output. On the application side, by optimizing the procedure to real-time, we can also explore the applications of this system in VR and AR for providing a more interactive experience.

REFERENCES

Barron, J. T., Adams, A., Shih, Y., and Hernández, C. (2015). Fast bilateral-space stereo for synthetic defocus. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4466–4474.

Barron, J. T. and Poole, B. (2016). The fast bilateral solver. In *European Conference on Computer Vision*, pages 617–632. Springer.

Blackmagic (2021). Ultimatte.

Bleyer, M. and Breiteneder, C. (2013). Stereo matching—state-of-the-art and research challenges. In *Advanced topics in computer vision*, pages 143–179. Springer.

Chuang, Y.-Y., Agarwala, A., Curless, B., Salesin, D. H., and Szeliski, R. (2002). Video matting of complex scenes. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*, volume 50, pages 243–248, New York, NY, USA. Association for Computing Machinery.

Chung, C.-Y. and Chen, H. H. (2009). Video object extraction via mrf-based contour tracking. *IEEE transactions on circuits and systems for video technology*, 20(1):149–155.

Fehrman, B. and McGough, J. (2014). Depth mapping using a low-cost camera array. In *2014 Southwest symposium on image analysis and interpretation*, pages 101–104. IEEE.

Ferstl, D., Reinbacher, C., Ranftl, R., Rüther, M., and Bischof, H. (2013). Image guided depth upsampling

using anisotropic total generalized variation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 993–1000.

Foundry (2021). Keylight-foundry learn. website.

Fusiello, A. and Irsara, L. (2008). Quasi-euclidean uncalibrated epipolar rectification. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. IEEE.

FXGuide (2014). The art of deep compositing. website.

Gvili, R., Kaplan, A., Ofek, E., and Yahav, G. (2003). Depth keying. In Woods, A. J., Merritt, J. O., Benton, S. A., and Bolas, M. T., editors, *Stereoscopic Displays and Virtual Reality Systems X*, volume 5006, pages 564–574. SPIE.

Heber, S., Ranftl, R., and Pock, T. (2013). Variational shape from light field. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 66–79. Springer.

Honegger, D., Sattler, T., and Pollefeys, M. (2017). Embedded real-time multi-baseline stereo. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5245–5250. IEEE.

Hosni, A., Rhemann, C., Bleyer, M., Rother, C., and Gelautz, M. (2012). Fast cost-volume filtering for visual correspondence and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(2):504–511.

Ihrke, I., Restrepo, J., and Mignard-Debise, L. (2016). Principles of light field imaging: Briefly revisiting 25 years of research. *IEEE Signal Processing Magazine*, 33(5):59–69.

Jeon, H.-G., Park, J., Choe, G., Park, J., Bok, Y., Tai, Y.-W., and So Kweon, I. (2015). Accurate depth map estimation from a lenslet light field camera. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1547–1555.

Kanade, Y., Yoshida, A., Oda, K., Kano, H., and Tanaka, M. (1996). A stereo machine for video-rate dense depth mapping and its new applications. In *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 196–202.

Kang, Y.-S., Lee, C., and Ho, Y.-S. (2008). An efficient rectification algorithm for multi-view images in parallel camera array. In *2008 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video*, pages 61–64. IEEE.

Klodt, M. and Vedaldi, A. (2018). Supervising the new with the old: learning sfm from sfm. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 698–713.

Kopf, J., Matzen, K., Alsisan, S., Quigley, O., Ge, F., Chong, Y., Patterson, J., Frahm, J.-M., Wu, S., Yu, M., et al. (2020). One shot 3d photography. *ACM Transactions on Graphics (TOG)*, 39(4):76–1.

Levin, A., Lischinski, D., and Weiss, Y. (2008). A closed-form solution to natural image matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):228–242.

Li, W., Viola, F., Starck, J., Brostow, G. J., and Campbell, N. D. (2016). Roto++ accelerating professional rotoscoping using shape manifolds. *ACM Transactions on Graphics (TOG)*, 35(4):1–15.

Li, Y., Sun, J., and Shum, H.-Y. (2005). Video object cut and paste. In *ACM SIGGRAPH 2005 Papers*, pages 595–600.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

Lu, T. and Li, S. (2012). Image matting with color and depth information. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 3787–3790. IEEE.

Luo, X., Huang, J.-B., Szeliski, R., Matzen, K., and Kopf, J. (2020). Consistent video depth estimation. *ACM Transactions on Graphics (TOG)*, 39(4):71–1.

Ma, F., Cavalheiro, G. V., and Karaman, S. (2019). Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 3288–3295. IEEE.

Manakov, A., Restrepo, J., Klehm, O., Hegedus, R., Eisemann, E., Seidel, H.-P., and Ihrke, I. (2013). A reconfigurable camera add-on for high dynamic range, multispectral, polarization, and light-field imaging. *ACM Transactions on Graphics*, 32(4):47–1.

Ming, Y., Meng, X., Fan, C., and Yu, H. (2021). Deep learning for monocular depth estimation: A review. *Neurocomputing*.

Nair, R., Ruhl, K., Lenzen, F., Meister, S., Schäfer, H., Garbe, C. S., Eisemann, M., Magnor, M., and Kondermann, D. (2013). A survey on time-of-flight stereo fusion. In *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*, pages 105–127. Springer.

Nozick, V. (2011). Multiple view image rectification. In *2011 1st International Symposium on Access Spaces (ISAS)*, pages 277–282. IEEE.

Ota, M., Fukushima, N., Yendo, T., Tanimoto, M., and Fujii, T. (2009). Rectification of pure translation 2d camera array. In *Proceedings of the Korean Society of broadcast engineers conference*, pages 659–663. The Korean Institute of Broadcast and Media Engineers.

Park, J., Kim, H., Tai, Y.-W., Brown, M. S., and Kweon, I. (2011). High quality depth map upsampling for 3d-tof cameras. In *2011 International Conference on Computer Vision*, pages 1623–1630. IEEE.

Park, K., Kim, S., and Sohn, K. (2018). High-precision depth estimation with the 3d lidar and stereo fusion. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2156–2163. IEEE.

Sara, U., Akter, M., and Uddin, M. S. (2019). Image quality assessment through fsim, ssim, mse and psnr—a comparative study. *Journal of Computer and Communications*, 7(3):8–18.

Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., and Westling, P. (2014). High-resolution stereo datasets with subpixel-accurate ground truth. In *German conference on pattern recognition*, pages 31–42. Springer.

Schonberger, J. L. and Frahm, J.-M. (2016). Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4104–4113.

Silberman, N. and Fergus, R. (2011). Indoor scene segmentation using a structured light sensor. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pages 601–608. IEEE.

Tao, M. W., Srinivasan, P. P., Malik, J., Rusinkiewicz, S., and Ramamoorthi, R. (2015). Depth from shading, defocus, and correspondence using light-field angular coherence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1940–1948.

Tao, T., Chen, Q., Feng, S., Hu, Y., and Zuo, C. (2018). Active depth estimation from defocus using a camera array. *Applied optics*, 57(18):4960–4967.

Vijayanarasimhan, S., Ricco, S., Schmid, C., Sukthankar, R., and Fragkiadaki, K. (2017). Sfm-net: Learning of structure and motion from video. *arXiv preprint arXiv:1704.07804*.

Wang, L., Gong, M., Zhang, C., Yang, R., Zhang, C., and Yang, Y.-H. (2012). Automatic real-time video matting using time-of-flight camera and multichannel poisson equations. *International journal of computer vision*, 97(1):104–121.

Wu, G., Masia, B., Jarabo, A., Zhang, Y., Wang, L., Dai, Q., Chai, T., and Liu, Y. (2017). Light field image processing: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 11(7):926–954.

Yang, J., Ding, Z., Guo, F., and Wang, H. (2014). Multi-view image rectification algorithm for parallel camera arrays. *Journal of Electronic Imaging*, 23(3):033001.

Yang, Q., Yang, R., Davis, J., and Nistér, D. (2007). Spatial-depth super resolution for range images. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.

Zhang, C. and Chen, T. (2004). A self-reconfigurable camera array. In *ACM SIGGRAPH 2004 Sketches*, page 151.

Zhou, C., Lin, S., and Nayar, S. (2009). Coded aperture pairs for depth from defocus. In *2009 IEEE 12th international conference on computer vision*, pages 325–332. IEEE.

Zhu, J., Wang, L., Yang, R., and Davis, J. (2008). Fusion of time-of-flight depth and stereo for high accuracy depth maps. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.