A Comprehensive Framework for Evaluation of Stereo Correspondence Solutions in Immersive Augmented and Virtual Realities

Bahar Pourazar[*], Oscar Meruvia-Pastor[‡]

[*]Computer Science Department
Memorial University of Newfoundland
Elizabeth Ave., St. John's, Canada
email: b.pourazar@mun.com
[‡]Computer Science Department
Memorial University of Newfoundland
Elizabeth Ave., St. John's, Canada
email: oscar@mun.ca

Abstract

In this article, a comprehensive approach for the evaluation of hardware and software solutions to support stereo vision and depth-dependent interactions based on the specific requirements of the human visual system within the context of augmented reality applications is presented. To evaluate stereo correspondence solutions in software, we present an evaluation model that integrates existing metrics of stereo correspondence algorithms with additional metrics that consider human factors that are relevant in the context of outdoor augmented reality systems. Our model provides modified metrics of stereoacuity, average outliers, disparity error, and processing time. These metrics have been modified to provide more relevant information with respect to the target application. We illustrate how this model can be used to evaluate two stereo correspondence methods: the OpenCV implementation of the semi-global block matching, also known as SGBM, which is a modified version of the semi-global matching by Hirschmüller; and ADCensusB, our implementation of ADCensus, by Mei et al.. To test these methods, we use a sample of fifty-two image pairs selected from the KITTI stereo dataset, which depicts many situations typical of outdoor scenery. Further on, we present an analysis of the effect and the trade-off of the post processing steps in the stereo algorithms between the accuracy of the results and performance. Experimental results show that our proposed model can provide a more detailed evaluation of both algorithms. To evaluate the hardware solutions, we use the characteristics of the human visual system as a baseline to characterize the state-of-the-art in equipment designed to support interactions within immersive augmented and virtual reality systems. The analysis suggests that current hardware developments have not yet reached the point where their characteristics adequately match the capabilities of the human visual system and serves as a reference point as to what are the desirable characteristics of such systems.

First presented at the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications VISIGRAPP 2015, extended and revised for JVRB

Keywords: Augmented Reality, Human Visual System, Binocular Stereo, Stereoacuity, Disparity, Stereo Correspondence, Virtual Reality, Field of View, Display Resolution, Depth Sensing Cameras, Head Mounted Displays

1 Introduction

Many Augmented Reality (AR) systems require some form of optical markers placed within a scene in order to integrate computer-generated objects with scenery directly generated from the real world; these markers help the system identify the location of an item within the scene to be used as a place-holder for the synthetic objects. Placing such markers in objects that are part of a scene may work for many indoor environments, but is a less practical option in outdoor AR settings where users can move freely in their surroundings. In the absence of such markers, an AR system requires a depth map of the surrounding environment. In order to obtain the 3D location of different objects in the scene, several technologies can be used. Among these technologies, one of the most practical techniques is the use of stereo cameras to take images of the scene from slightly different viewpoints. These images can then be processed by the stereo correspondence algorithms, which attempt to find the corresponding pixels in the stereo images, to generate the depth map of the surrounding environment. This map is then used to integrate virtual objects in the scene such that synthetic objects are rendered in a way that considers the occlusion properties and the depth of the real objects in the scene. Due to the potential applications of stereo correspondence, which is one of the most extensively studied subjects in computer vision [SS02], using an evaluation scheme that is designed according to the specific requirements of the target application is essential. The evaluation scheme proposed in this paper is designed for outdoor AR applications which make use of stereo vision techniques to obtain a depth map of the surrounding environment.

Over the past few years, a few evaluation schemes have been proposed by researchers in the field to provide a testbed for assessment of the solutions based on specific criteria. The Middlebury Stereo [SB12] and the KITTI Stereo benchmarks [GLU12] are two of the most popular and widely used evaluation systems through which a stereo correspondence algorithm can be evaluated and compared to others and each year they call out for new submissions of stereo correspondence algorithms by researchers in the community to update their evaluation results. In spite of being a valid reference in many applications, both the KITTI and the Middlebury projects take a general approach towards evaluating the methods; that is, they have not been designed with an eye to any particular target application. In fact, these models focus on the particular application of a stereo correspondence algorithm as a solution per se to find the *best matches* of the corresponding pixels in stereo pairs, regardless of the target application. As a result, the information provided by these evaluation benchmarks is not sufficient to select a given algorithm suitable for AR because we need information on the specific *accuracy* and *efficiency* of these algorithms, for example, to assess their suitability regarding their processing time or accuracy. The fact that some of this information is missing from such standard evaluations of stereo correspondence methods has compelled us to take steps towards a comprehensive analysis and evaluation design based on specific requirements of outdoor stereo AR applications, which results in better definition and adjustment of the criteria for efficiency and accuracy metrics used for the evaluation.

2 Background and Terminology

Over the past decades, many mobile AR systems have been built, from the Touring Machine in 1997 [FMHW97] to Google Glass which was announced in 2013 [Goo13]; however, most of these prototypes have remained experimental due to certain difficulties and constraints of using them in practical applications [DM96, Liv05]. Two of the most important constraints are the human factors in AR and the high demand of computational resources needed to provide a real-time interaction between the user and the system; therefore, in order to build a practical AR system, these factors need to be carefully considered while designing different components of the system.

2.1 Binocular Vision and the Human Visual System

Studies in binocular vision show that human perception of depth can vary depending on the environment and under different circumstances. Many studies have focused on the evaluation of human perception of depth within different frameworks and in different applications, such as virtual reality and AR, which have recently attracted more attention [WRMW95, DM96, Liv05, JW05, SJK$^+$07, KSF10, SSJE12, DS14]. These studies show that the viewer perception of depth is inversely proportional to his/her distance from the object [KSF10, SJK$^+$07, JW05, Liv05]; for instance, in [SJK$^+$07], some experiments are designed to study and evaluate the human perception of distance, in terms of the absolute depth of the objects from the observer, for an outdoor AR application in urban settings. However, in this research we are more interested in the human perception of relative depth in stereo vision: the ability to perceive and distinguish the depth of different objects relative to each other. Binocular disparity, which in fact arises from the spatial difference between the images of the same scene in the visual system, provides a relative perception of depth from the surrounding environment. This perception is known as *binocular stereopsis* [HR95]. In stereo vision, the locus of the points that yield a unified view of an object in the visual system is known as the *horopter*, and any point located on the horopter is usually called a *fixation point* [Rea83, HR95]. An important property of an object on the horopter is that no spatial difference exists between the images of the fixated object between the two eyes, that is, the binocular disparity is zero [HR95]. Exploiting this property, the disparity of any other object in the scene can be estimated relative to the fixated object by inspecting two important factors: whether the object of interest is closer or further than the fixated object and then how much closer or further it is relative to the fixated object. As a result, the binocular disparity provides a relative perception of depth of the surrounding environment. In binocular vision, the minimum depth difference between two points that can be detected in the visual system is known as *Stereoscopic Acuity* or *Stereoacuity* [Pfa00]. This metric is normally presented in angular units, commonly arcseconds. According to the geometry of binocular vision illustrated in Figure 1, stereoacuity can be obtained from the following equation:

$$\theta = \frac{a\Delta Z}{Z^2} \tag{1}$$

This equation estimates the angular disparity in radians, where a is the distance between the center of the pupils of the two eyes, which is known as interpupillary distance.

According to standard stereo tests [Rea83], the finest detectable disparity in the human visual system (HVS) is approximately 10-15 arcseconds. However, a more recent study on 60 subjects [GS06] at different age groups, from 17 to 83 using standard stereotests, shows that the average stereoacuity for different age groups is as follows:

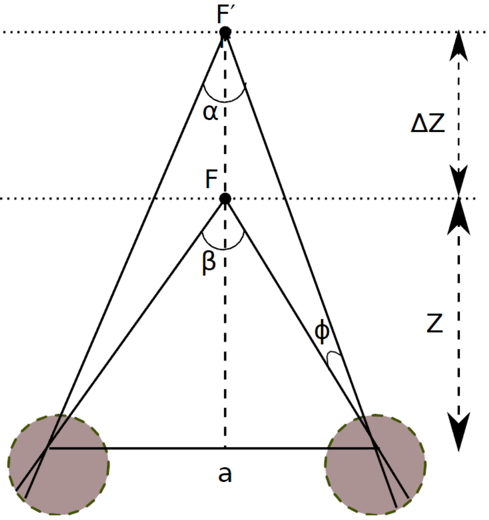


Figure 1: Binocular disparity

Age Range	Avg_Stereoacuity(arcsecs)
17-29	32
30-49	33.75
50-69	38.75
70-83	112.5

Table 1: Average stereoacuity for subjects of age 17 to 83

As can be seen, the stereoacuity for the HVS increases with age, that is the amount of error in the depth results is less perceptible in the visual system of the elders than the younger people. Using these values in Equation 1 along with the average interpupillary distance in the human visual system, that is approximately $64mm$ [HR95], we can estimate the threshold for minimum detectable depth between two objects based on their distance from the observer.

2.2 Real-time Interaction

Providing real-time interaction in an AR system for the user requires the processing time and update rate of the whole system to keep up ideally with the standard video frame rate, between 24fps and 30fps, or higher. However, studies show that in practice to build a reasonable interactive augmented world the processing rate should not be less than half of the video frame rate [HP00]. Two ways to speed up a system are using a more advanced technology and hardware, and implementing more sophisticated and efficient software design. However, having access to advanced technology and hardware is not always feasible and even the most advanced technologies have some limitation in their memory space and computational capability which may not meet the requirement for some real-time applications. Therefore, in many cases, employing the second approach is more practical.

3 Comprehensive Evaluation Scheme

In our design, unlike the Middlebury or KITTI benchmarks, we label a pixel in the disparity results as an *outlier* if the angular measurement, that is the stereoacuity, corresponding to the depth error between the ground truth and the estimated depth value by the algorithm is more than the average perceptible stereoacuity of the HVS as determined by standard stereo tests [Rea83, GS06]. Moreover, we use the average stereoacuity for different age groups [GS06] in our design to evaluate the performance of the algorithm for users at different ages; this makes the evaluation results more reliable and applicable to applications of AR. In order to evaluate the efficiency of an algorithm and investigate whether it meets the requirements for being part of a real-time AR application, we integrate a module in the evaluation process that reports on the average execution time of the algorithm for the input data. The average outliers based on the specified stereoacuity thresholds and the average disparity error are also estimated during the evaluation process.

In addition, our model employs a particular approach which can be of specific value to AR applications. In this approach, we suggest that it is important to focus the evaluation process on particular regions of the disparity map rather than the whole image. The main reason is that salient edges caused by depth discontinuities, which also represent object boundaries and occlusion, are important depth cues for the human visual system to better perceive the location of different objects in the 3D environment [Sze11]. Therefore, more accurate depth results in these regions permit a higher quality combination of the depth map of the real world with the virtual depth of the synthetic objects that are part of the AR scene. To this end, we build a mask of the ground truth disparity map by applying the Canny edge detector and then the Dilation operation to include areas from both background and foreground of the scene near the edges for our evaluation. The result mask is, in fact, a mask of the edges in the image caused by depth discontinuities and their surrounding area.

3.1 Architecture

The block diagram of our evaluation system can be seen in Figure 2, which illustrates the sequence of the operations during the whole process. As can be seen in this figure, first the input data consisting of the stereo images, the ground truth disparity, and the calibration data are passed to the system. Afterwards, the specified masks are created using a *Canny* edge detector and a *Dilation* operation with the appropriate parameters selected separately for each image. After the corresponding disparity maps have been generated by the stereo algorithm and stored on the disk, they are passed to the evaluation module with the specified arguments. Finally, the evaluation metrics are estimated and output as data files and plots to facilitate the evaluation of the stereo algorithm in the application of interest.

3.2 Evaluation Metrics

The main evaluation component consists of different modules which estimate specific evaluation metrics. These metrics are: 1) the average stereoacuity, 2) the average outliers, 3) the average disparity error, and 4) the average execution time. Analysis of these metrics in the framework of an outdoor AR application will then allow for a practical evaluation of the stereo algorithm performance.

3.2.1 Average Stereoacuity

We can break the estimation of the average stereoacuity down to 3 steps: 1) estimate the stereoacuity based on the generated disparity for each image pair and the ground truth; 2) average the stereoacuity results over certain depth ranges in each image; 3) average the results from the previous step over all the images. Corresponding plots are generated after the third step based on the final results.

According to the specific age ranges, different values are reported for the average stereoacuity at the end of the evaluation. In order to estimate this metric, the depth values corresponding to both ground truth and the generated disparity by the algorithm are first calculated. Subsequently, the difference between these values is used in Equation 1 to calculate the corresponding stereoacuity, Equations 2 and 3. This process is done for all the pixels in the image; or if a mask has been provided, it will be applied only to the pixels in the masked areas. Finally the results are output and

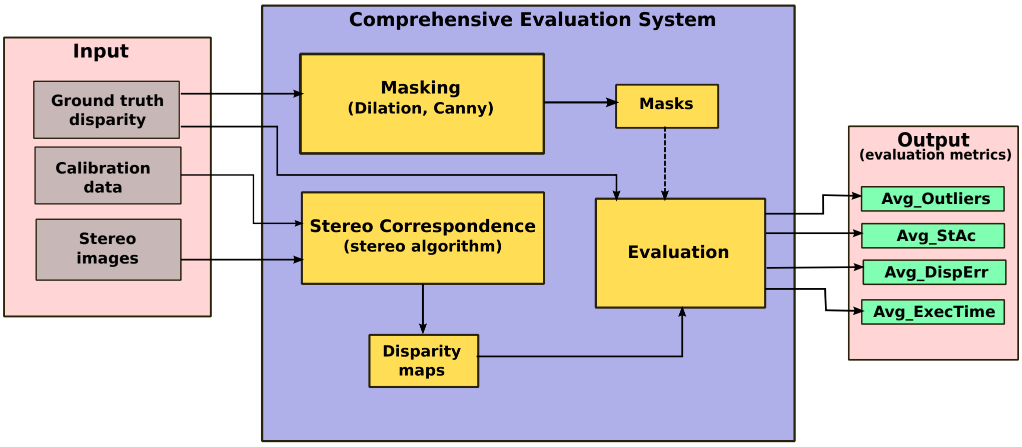


Figure 2: Architecture of the evaluation system

stored in a separate data file for each image.

$$Depth_{err} = |depth_{gt} - depth_{gen}| \tag{2}$$

$$StAc = \frac{a * Depth_{err}}{depth_{gt}^2} \tag{3}$$

Here, $depth_{gt}$, $depth_{gen}$, $StAc$ and a denote the ground truth depth, the generated depth by the algorithm, the corresponding stereoacuity, and the average interpupillary distance, respectively.

After conducting the first step on all the disparity maps corresponding to input image pairs, the second step starts by building a histogram of the stereoacuity values over specific depth ranges, Equations 5 and 6. In our design, the width of each bin determines the aforementioned depth range and is kept constant for all the bins. Moreover, the number of bins along with their corresponding width determine the total distance over which the results are estimated and subsequently examined.

$$Total\_distance = NumOfBins * Width \tag{4}$$

For outdoor applications of AR, these parameters are normally set to certain values so that the total distance covers the medium to far depth fields; extending from 1.5 meters to more than 30 meters [SJK+07].

$$Sum_{dRange} = \sum_{dRange} StAc \tag{5}$$

$$Avg\_StAc_{dRange} = \frac{Sum_{dRange}}{NumOfPixs_{dRange}} \tag{6}$$

Here, Avg_StAc_{dRange} and Sum_{dRange} denote the average and total stereoacuity over specified depth ranges in each image, and $NumOfPixs_{dRange}$ denote the number of pixels within each depth range.

The results of the previous step, all stored in a data file, are then passed to the last step. At this point, a histogram is built over the data from all the disparity images, which results in the average stereoacuity values within each specified depth range over all the images, Equation 7. It should be noted that the number of bins and their corresponding width in this step are similar to the histogram constructed in the previous step.

$$Avg\_StAc = \frac{\sum_{imgs}(Sum_{dRange})}{\sum_{imgs}(NumOfPixs_{dRange})} \tag{7}$$

3.2.2 Average Outliers

For this measurement, the relative depth error is first calculated by finding the corresponding depth values for the ground truth disparity and the disparity generated by the algorithm and then converted to effective stereoacuity, as shown in Equations 2 and 3. This value is then compared to the relative detectable depth threshold for the HVS that is estimated using Equation 1. If the relative depth error is equal to or more than the detectable threshold in the HVS, Equation 8, then the corresponding pixel is labelled as an outlier.

$$StAc \geq StAc_{threshold} \tag{8}$$

Since we are using four different thresholds of stereoacuity corresponding to different age groups in our evaluation, the estimated error is compared against each of these thresholds, and therefore, four different values are eventually calculated. The average outliers is then computed as a fraction of the total number of pixels in the inspected regions, Equation 9.

$$Avg\_Outliers = \frac{Outliers_{total}}{NumOfPixs} \tag{9}$$

This process is repeated for all the pixels in the image or merely the pixels in the masked regions depending on the availability of a mask.

3.2.3 Average Disparity Error

This metric is the mean error between the ground truth disparity and the one found by the algorithm, which is estimated for all the pixels in the image or merely the masked pixels depending on the availability of a mask. It can be presented with the following estimations:

$$Disp_{err} = |disp_{gt} - disp_{gen}| \tag{10}$$

$$DispErr_{total} = \sum_{pixs} Disp_{err} \tag{11}$$

After the computation of the total disparity error for the pixels, the average disparity error is estimated as follows:

$$Avg\_DispErr = \frac{DispErr_{total}}{NumOfPixs} \tag{12}$$

The *NumOfPixs* is, in fact, the total number of pixels in the whole image or the masked regions, depending on the case for which the error is being estimated.

3.2.4 Average Execution Time

We use the C++ function *clock()* to estimate the average execution time of the algorithms for generating disparity results corresponding to the input stereo images, fifty-two image pairs in our evaluation. We then compare this value to the acceptable criteria for having a real-time interactive AR system from the user's perspective, that is, a processing time less than 0.06-0.08 seconds per frame corresponding to a frame rate of 12.5 to 16.5 fps, as proposed by [HP00].

Analyzing each of these metrics in the light of the relevant factors in an outdoor AR application results in a practical evaluation of the stereo correspondence methods.

4 Validation

In order to verify the effectiveness of our proposed model for the evaluation of stereo correspondence methods in outdoor AR applications, we have evaluated two sample stereo algorithms: the OpenCV implementation of the semi-global block matching, also known as SGBM, which is a modified version of the semi-global matching by Hirschmüller [Hir08]; and ADCensusB, our implementation of "on building an accurate stereo matching system on graphics hardware" [MSZ$^+$11], originally known as ADCensus. It should be noted that the CPU implementation of both methods have been used.

Experiments were carried out on a Linux platform with Intel Core i7 3.20GHz CPU. Fifty-two image pairs were chosen from the KITTI Stereo Dataset corresponding to real outdoor scenes. Figure 13 shows a sample stereo pair from the KITTI dataset. The OpenCV Canny edge detector and Dilation operation were used for building the specified masks and the expansion of the masked areas, respectively. The corresponding mask of Figure 13, is shown in Figure 14. The masked ground truth and the masked disparity images generated by SGBM and ADCensusB for the sample stereo image, can also be seen in Figures 15, 16 and 17, respectively.

Parameters corresponding to stereo algorithms, the aperture size in Canny, and the degree of Dilation were kept constant over all the images and experiments. These values are presented in Tables 2, 3, and 4. The parameters for Dilation and Canny were chosen empirically by running the algorithms over our image set with the intention of selecting the values which best define the depth edges and expand them enough to include regions with different depths surrounding the edges.

SADWindowSize	9	disp12MaxDiff	2
uniquenessRatio	10	P2	3*9
speckleWindowSize	100	speckleRange	2

Table 2: SGBM Parameters

The minimum and maximum disparity values are also kept constant for each image pair in both algorithms; however, the maximum disparity differs for each image pair as the scenes are different and objects are located at different depth fields. The minimum disparity is set to 0 for both algorithms. The maximum disparity for each image pair is selected based on the maximum value in their corresponding ground

λ_{AD}	10	λ_{Census}	30	L_1	34	L_2	17
τ_1	20	τ_2	6	π_1	1.0	π_2	3.0
τ_{SO}	15	τ_S	20	τ_H	0.4		

Table 3: ADCensusB Parameters

Dilation_iterations	10
Canny_apertureSize	3

Table 4: Masking Parameters

truth disparity. The standard stereoacuities used for the evaluation are based on the results mentioned in Table 1.

4.1 Experimental Results

The evaluation metrics, mentioned in Section 3.2, were estimated for SGBM and ADCensusB in our evaluation system. The main results are described below.

4.1.1 Average Stereoacuity

Figures 3 and 4 show the average relative depth error converted to effective stereoacuity over distance for the masked and the whole images with both SGBM and ADCensusB.

In these plots, a cross point below a stereoacuity threshold (straight lines) implies that the average error in the disparity values estimated by the stereo algorithm is imperceptible to the human visual system. However, a value higher than the threshold indicates that the error cannot be ignored and should be resolved to achieve a better alignment between the virtual and the real world in the AR application of interest. Moreover, as can be seen most of the errors fall below the standard stereoacuity value corresponding to older ages; indicating that these are not perceptible to the visual system of people at these particular ages.

The zero values in the plots imply that either there is no object within the corresponding range or the disparity value estimated by the algorithm is equal to the ground truth disparity; however, since the average of the results has been taken over all the images, it is more likely that the zero values indicate no object within the particular range.

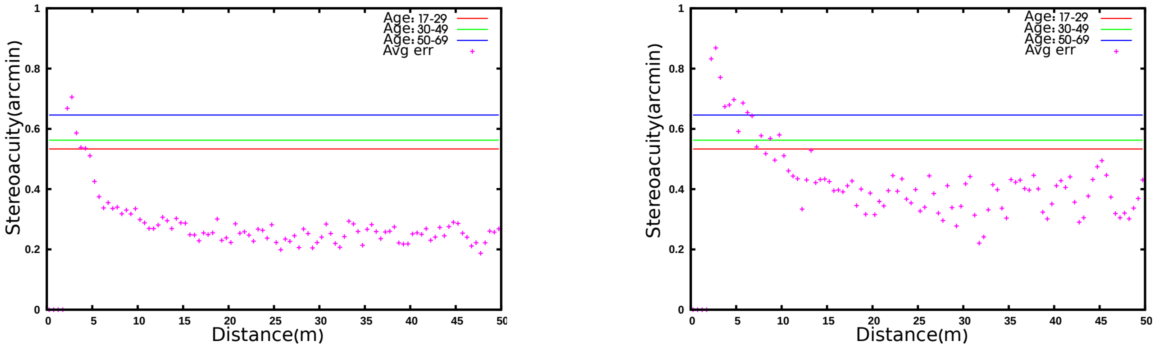
As can be seen in the results, SGBM performs better in finding more accurate corresponding matches compared to ADCensusB, as most of the error points fall below the standard stereoacuity lines. Moreover, the plots show that in both methods the significant amount of error corresponds to the near field objects, within the first 5 meters. This range of the depth field can be considerably important in some applications, such as the ones involving certain manipulative tasks; for these types of applications, other technologies, such as depth sensing cameras, are better choice.

Comparing the results between the masked and the whole image show that the average error over the masked regions, that is, near the depth edges, is very similar to the results over the whole image. This may imply that there is no additional benefit in the inspection of these regions. However, this might be merely an indication of the performance of the selected algorithms and can be better analyzed by evaluating more algorithms within our model.

4.1.2 Average Outliers

The average outliers for the masked and the whole image are presented in Table 5. Results show that in both cases, the masked regions and the whole image, SGBM has less average outliers than ADCensusB, indicating that SGBM generates a more accurate disparity map as perceived by the human visual system.

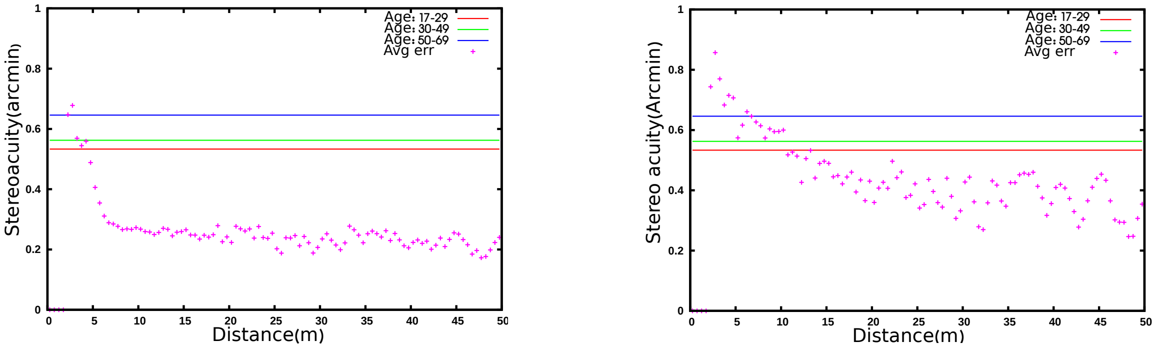
Another observation is that in SGBM, the average outliers over the masked regions is larger than the average outliers over the whole image, whereas in ADCensusB the opposite behavior is observed. This implies that SGBM generates less accurate results near the depth discontinuities and occluded regions compared to the other areas in the image. On the other hand, ADCensusB generates more accurate disparity values near the depth edges compared to the other regions in the image and tends to preserve the occluded regions. This only indicates that, despite the better performance of SGBM over ADCensusB according to the experimental results, in cases where only one of these solutions is available, it is reasonable to consider this behavior to employ the method in the right application based on the accuracy requirement of the target system in different regions. In other words, it is important to first investigate which regions of the image are more important in the context of the target application. For instance, ADCensusB performs better in an application where the areas near depth discontinuities and occlusion are more important than the rest of the image, such as image compositing for layering visual elements on the scene, compared to application scenarios where obtaining an accurate, dense disparity map for all the regions in an image is essential, such as constructing a 3D model of the scene or preparing a model for 3D printing. Figure 11 shows a comparison of all the results.



(a) Average relative depth error by SGBM

(b) Average relative depth error by ADCensusB

Figure 3: Average relative depth error over distance for the masked image



(a) Average relative depth error by SGBM

(b) Average relative depth error by ADCensusB

Figure 4: Average relative depth error over distance for the whole image

Algorithm	Age	Avg_Outliers (Masked)	Avg_Outliers (Full)
SGBM	17-29	0.12	0.11
	30-49	0.11	0.10
	50-69	0.09	0.08
	70-83	0.0012	0.005
ADCensusB	17-29	0.23	0.27
	30-49	0.22	0.26
	50-69	0.18	0.22
	70-83	0.002	0.002

Table 5: Average outliers

Algorithm	Region	Avg_DispErr
SGBM	Full	6.58
	Masked	7.81
ADCensusB	Full	4.49
	Masked	4.74

Table 6: Average disparity error

4.1.3 Average Disparity Error

The average disparity error for both the whole and the masked image are presented in Table 6. As can be seen, ADCensusB results in less average disparity error than SGBM. This difference is likely caused by the various refinement steps implemented in the ADCensusB algorithm which do not exist in SGBM. As a result, despite the larger outliers in ADCensusB than SGBM as presented in Section 4.1.2, ADCensusB attempts to decrease the difference between the resulting disparity value and the ground truth disparity through multiple refinement steps, thus generating smoother disparity patches within different regions of the image. Figure 12 presents a comparison between all the results.

4.1.4 Average Execution Time

In another experiment, we estimated the average execution time for both algorithms using a set of fifty-two stereo image pairs from the KITTI data set [GLU12]. Results of the average execution time over all the images are shown in Table 7. Considering the requirements of a real-time AR system [HP00], the processing time of each frame should not be more than 0.06-0.08 seconds. Although the current implementation of SGBM could be used when the real world scene remains stable for approximately one second, it can be safely concluded that none of these implementation meets the requirements of a real-time interactive AR system.

Algorithm	Avg_ExecTime(secs)
SGBM	0.54
ADCensusB	272.82

Table 7: Average execution time

4.2 Effect of Refinement

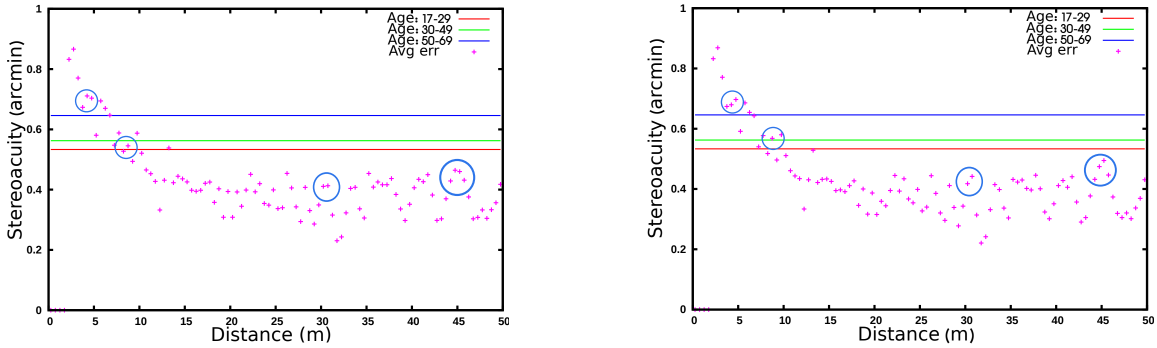
In this experiment, we studied the effect of the post processing steps, also referred to as the *refinement steps*, in the stereo algorithms on the accuracy of the results in our evaluation criteria.

Refinement is usually the last step in a stereo correspondence algorithm because it attempts to decrease the number of wrong matches or the error after the disparity results have been found [SS02]. Therefore, this step must be applied after the outliers, that is the wrong pixel matches, have been detected in the results. The detection of the outliers occurs through a check known as left-right consistency check in a stereo matching algorithm. In this check, the disparity map for both the left and right image is first calculated. Then, if a pixel in the left image, based on its disparity value, corresponds to a pixel in the right image that does not map back to it, it will be labeled as an outlier [SS02]. This description can be formulated as follows:

$$D_L(p) \neq D_R(p - (D_L, 0)) \tag{13}$$

Where $D_L(p)$ is the disparity function for the left image and D_R is the disparity function for the right image.

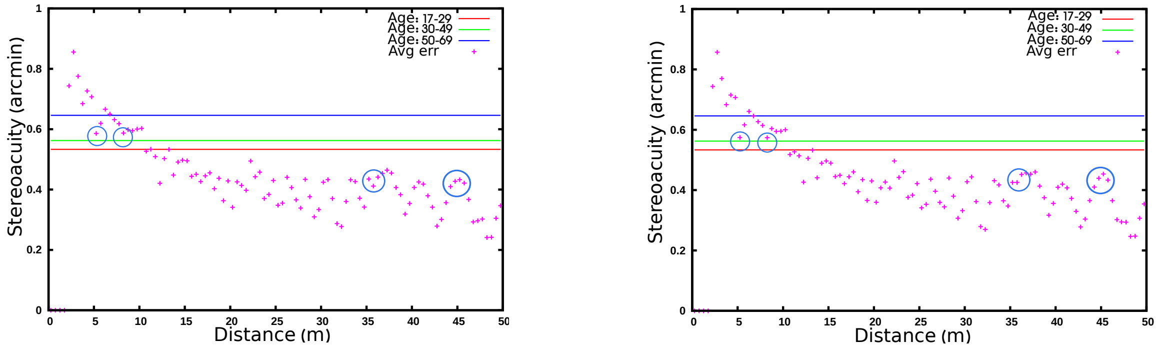
In our implementation of ADCensusB, we have the L-R check and its subsequent refinement steps triggered with a flag. Therefore, when the flag is not set, neither the check nor the refining steps are triggered in the algorithm. To investigate the effect of the refinement on the final results, we used ADCensusB in this experiment with the L-R flag set to zero, generating the disparity results for the image pairs, and evaluating the results. The results for both cases, not refined



(a) No refinement on disparity results

(b) Refinement on disparity results

Figure 5: Average disparity error by ADCensusB for the masked images; blue circles show some sample values that have slightly changed as a result of refinement



(a) No refinement on disparity results

(b) Refinement on disparity results

Figure 6: Average disparity error by ADCensusB for the whole images; blue circles show some sample values that have slightly changed as a result of refinement

Region	Avg_DispErr
Masked	5.59
Full	5.29

Table 8: ADCensusB average disparity error - unrefined

		Avg_Outliers	
Region	Age	valid_gtDisp	valid_genDisp
Masked	17-29	0.23	0.33
	30-49	0.22	0.31
	50-69	0.18	0.27
	70-83	0.002	0.003
Full	17-29	0.27	0.39
	30-49	0.26	0.37
	50-69	0.23	0.32
	70-83	0.001	0.002

Table 9: ADCensusB average outliers - unrefined

ADCensusB	Avg_ExecTime (secs)
refined	272.82
unrefined	147.84

Table 10: ADCensusB average execution time - refined and unrefined

and refined, over the masked regions and the whole image are shown in Figures 5 and 6, respectively.

As can be observed in the plots of Figures 5 and 6, the evaluation results in our specific criteria are not significantly different from the results of the algorithm when L-R check and refinement were triggered and only a few average values have slightly changed. We have marked a few of these values with blue circles in Figures 5, 6.

We also estimated the average execution time, the average disparity error, and the average outliers in this experiment. The results for the average error and outliers are shown in the tables below.

Figure 7 shows a comparison between the average outliers by ADCensusB with the effect of refinement and without it for the masked and the whole images in one of the validity criteria, that is, when the ground truth disparity is valid. As can be seen, no significant decrease is obtained in the number of outliers.

The average execution time was approximately 147.84 seconds which is nearly half the running time of the algorithm with the L-R check and refinements triggered, Table 10. Comparing these results to the ones presented in Tables 5 and 6 a slight decrease in the amount of errors and nearly no change in the number of outliers is observed. Analyzing the results in this experiment, we can conclude that despite the considerable rise in the execution time of the algorithm, no significant improvement in accuracy is achieved in our evaluation criteria through refinement of the disparity results; therefore, the execution of ADCensusB without any L-R check and refinement step is more beneficial to an AR application in outdoor environments, since it requires less processing time with a modest decrease in quality.

4.3 Discretization Degree of Disparity Values

According to different studies [DM96, ABB+01, KSF10, Cor12], some other factors such as issues associated with the environment, display device, and capturing device can also affect the perception of depth in the visual system. As a result, the ability to detect the difference in depth and to accurately estimate the depth of different points, in practice, do not merely depend on the implemented discretization level of the disparity values in the stereo correspondence algorithm. In order to investigate the validity of this statement, we conducted the following experiment. In this experiment we defined some stereoacuity thresholds. In order to find the minimum threshold to start with, we attempted to find the minimum disparity change in the ground truth disparity images. To this end, we move along horizontal scanlines in each image and compute the difference between the values of consecutive pixels, which is, in fact, an indicator of the detectable threshold of the changes in depth between different pixels. This is illustrated in Figure 8.

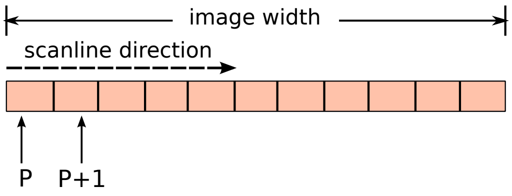


Figure 8: The scanline pixels difference process

After finding the minimum value in each image, a global minimum is sought between all the computed values from different images. The value we found for

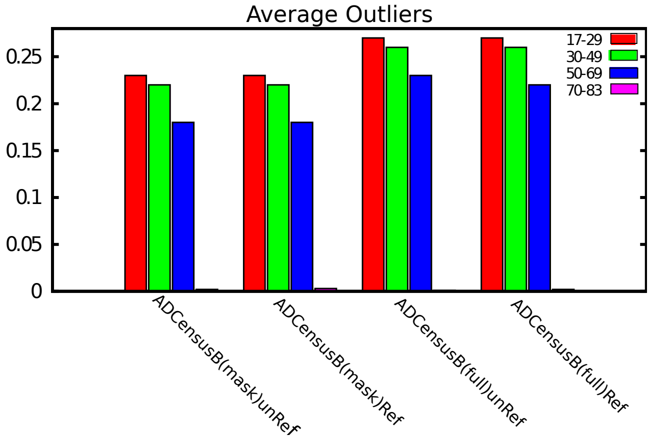


Figure 7: Average outliers by ADCensusB in refined and unrefined cases for both masked and whole images; each bin color corresponds to different age groups with specific stereoacuity thresholds

a group of twelve images selected from our dataset with the size of 1242×375 and the focal length of 721 pixels, as reported by KITTI stereo [GLU12], was 0.0022 arcmins. After finding this minimum and defining our thresholds, we apply a nearly similar operation on the disparity results of the same group of images from each of the sample algorithms. In this process, while moving through each image, for those pixels whose generated disparity is close to the ground truth disparity, within a specific pixel threshold, we estimate their depth difference from their following pixel in the ground truth and compare the value to each of the specified thresholds; if this value is less than a threshold, then we check to see whether the stereo algorithm has also detected different values for the corresponding pixels. In case of detection, we increment a counter corresponding to each threshold that indicates the number of detected pixels. This process is repeated for different images and, finally, the average of detected pixels is estimated for each specified threshold.

The results for both algorithms are shown in Figure 9. As can be seen in these plots, for both algorithms, the average detected pixels with detectable change in depth values starts to converge at the value of approximately 0.4 arcmins. We also observe that for the values below this threshold, the average detected pixels are very small and for some values, such as the minimum detectable threshold in ground truth, both algorithms are not capable of detecting any change in depth values. This implies that, regardless of the accuracy resolution of the algorithms, which is $\frac{1}{8}$th of a pixel for SGBM, approximately 0.6 arcmins, and $\frac{1}{16}$th of a pixel for ADCensusB, approximately 0.3 arcmins, for KITTI images based on the camera parameters and the geometrical relation presented in Figure 10 and Equation 14, some changes in depth in the real world still cannot be detected by the algorithm. This effect might be due to the constraints of the sensor, that is, the errors associated with the capturing device and its resolution, or the environmental noise.

In Figure 10, w is the image width and f is the focal length of the capturing device.

$$\theta = \arctan(\frac{pixel\_resolution}{focal\_length}) \tag{14}$$

For the image size of 1242×375 pixels and focal length of 721 pixels, and based on the resolution of SGBM and ADCensusB in the estimation of the disparity values, the minimum and maximum detectable disparity, that is, at the center and at the boundary of the image, respectively, in terms of effective stereoacuity are as follows:

$$\begin{aligned} &SGBM:\\ \theta_{max} &= \arctan(\frac{\frac{1}{8}}{721})\\ &= 0.00993 \text{ degrees} \times 60\frac{arcmins}{degrees}\\ &= 0.596 \text{ arcmins} \end{aligned} \tag{15}$$

$$\begin{aligned} & SGBM: \\ \theta_{min} &= \arctan(\frac{\frac{1}{8} \times (\frac{1242}{2})}{721}) \\ &- \arctan(\frac{\frac{1}{8} \times (\frac{1242}{2} - 1)}{721}) \\ &= 0.589 \text{ arcmins} \end{aligned} \tag{16}$$

$$\begin{aligned} & ADCensusB: \\ \theta_{max} &= \arctan(\frac{\frac{1}{16}}{721}) \\ &= 0.00496 \text{ degrees} \times 60 \\ &= 0.298 \text{ arcmins} \end{aligned} \tag{17}$$

$$\begin{aligned} & ADCensusB: \\ \theta_{min} &= \arctan(\frac{\frac{1}{16} \times (\frac{1242}{2})}{721}) \\ &- \arctan(\frac{\frac{1}{16} \times (\frac{1242}{2} - 1)}{721}) \\ &= 0.297 \text{ arcmins} \end{aligned} \tag{18}$$

However, as can be seen, the minimum and maximum angular resolution in the image are not considerably different.

As a result of this experiment, we can conclude that in order to achieve more accurate depth results in the stereo algorithms and correctly detect the difference between depth values, that is, to obtain a lower threshold of depth changes closer to the actual resolution of the implemented algorithm, using higher resolution devices and considering their robustness to noise is also essential. Based on the information about the average stereoacuity in the HVS, we can say that the lower bound resolution of a capturing device with focal length of 721 pixels should be $\frac{1}{8}th$ of a pixel. In the end, we should note that the experimental results presented earlier in this research show that despite various types of errors relevant to the capturing device, environmental noise, and the actual accuracy of the stereo correspondence algorithm in the estimation of disparity values, the effect of such errors on the results will still be imperceptible for most cases to the HVS in outdoor AR applications, especially where objects are distant from the observer.

Metrics	**Evaluation Models**		
	Middlebury	*Kitti*	*Comprehensive Evaluation*
Avg_StAc	✗	✗	✓
Avg_Outliers	✓	✓	✓
Avg_DispErr	✓	✓	✓
Avg_ExecTime	✗	✗	✓

Table 11: Comparison of different evaluation schemes

4.4 Overview

Table 11 shows an overview of the difference between our proposed evaluation approach and the other evaluation models, Middlebury and KITTI, in terms of the estimated evaluation metrics.

It should be noted that although the average error and the average outliers exist in the other evaluation schemes as well, the major difference which makes our evaluation more appropriate than the other schemes for practical applications of AR, is the approach employed during the design of the metrics and the analysis of the results in the evaluation process. In fact, integrating the important factors related to the human visual system and its perception of depth in the design of the metrics and the insights they provide make the evaluation model more relevant and applicable to outdoor AR systems.

5 Metrics for Immersive Augmented and Virtual Realities Equipment

In this section we present a detailed analysis of essential equipment to support stereoscopic vision and depth-sensing capture of hand gestures in the user's surroundings in Immersive Realities (AR/VR) systems. The analysis starts from the basis that the Human Visual System should be used as a reference to define the desired targets for the amount of visual detail that should be captured and displayed, the field of view of cameras and displays, and the range of action the depth-sensing cameras should support to provide comprehensive support for optimal immersive AR & VR interactions.

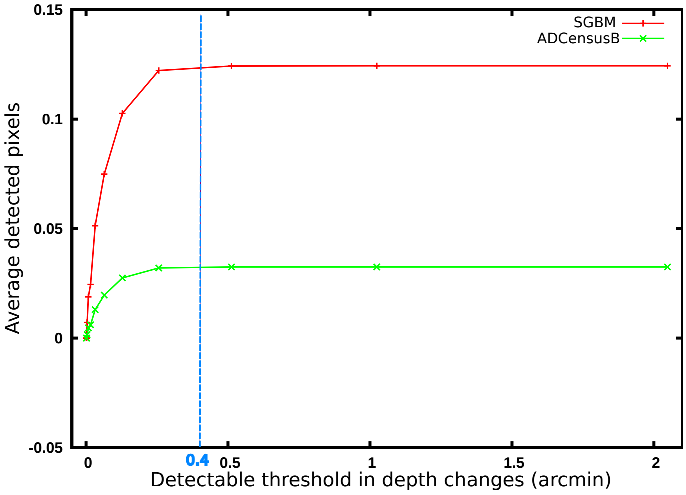


Figure 9: Average of detected pixels by SGBM and ADCensusB for specific stereoacuity thresholds marked on each curve for a group of 12 images; the vertical blue line shows the approximate threshold after which the average of detected pixels converge

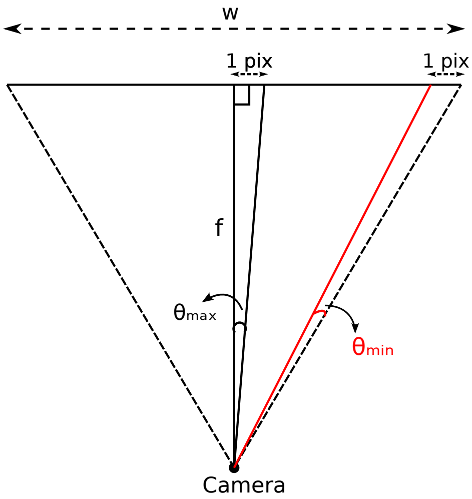


Figure 10: Resolution of image in angular disparity

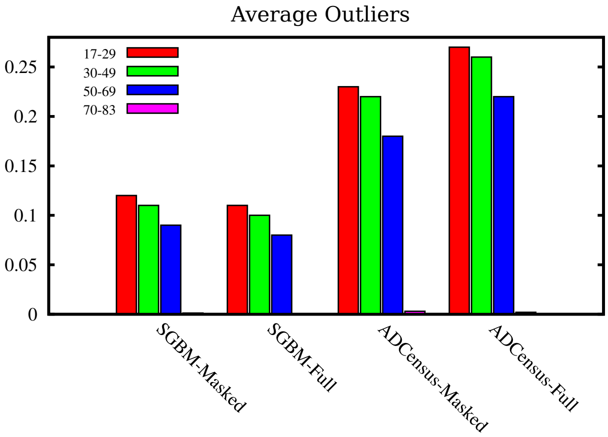


Figure 11: Average outliers for both SGBM and ADCensusB

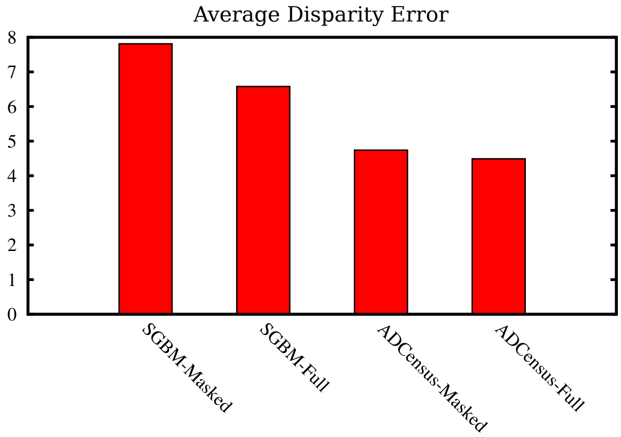


Figure 12: Average disparity error for both SGBM and ADCensusB

5.1 Screen Resolution and Sensor Density in the Retina

Humans have about 125 million light sensors in their retinas [Hec98]. If these sensors were mapped to a square grid, this would correspond to approximately 11,200 sensors on the horizontal and vertical sensors. If the sensors of the retina were mapped to a circle, which would be more fitting shape to represent the retina, the circle would have a radius of 6,300 sensors. These figures can be used as the baseline to determine the approximate resolution a display should have to match the number of sensors in the eye. For the sake of simplicity, we will assume that one pixel in a display may roughly correspond to one sensor in the retina. Since the human eye has a field of view of nearly 180 degrees, an initial estimate on the number of sensors required to replicate an impression in each of the retinas would result in about 70 samples (pixels) per degree, based on a circular mapping with a diameter of 12,600 sensors for 180 degrees. Accordingly, we propose setting the baseline of display resolution to 70 pixels per degree (ppd) to support peripheral vision. A similar principle was used when Apple introduced its line of retina displays. Apple's retina displays present in different products such as the Apple Watch, iPad and iMac Retina 5K, range from 57.5 ppd to 86 ppd [Wik15]. It is worth noting that the physical size of the display, its resolution and typical distance of use are taken into account to estimate these figures. Apple's ppd estimates are based on the capacity of displays to support 20/20 vision, a relative measure of vision describing an individual who can see as well as most other people.

Not all light sensors in the retina work the same, some work as illumination sensors (rods) and some work as color sensors (cones). The cones are divided in three color sensor types, specialized in the perception of each of the red, green, and blue light frequencies. In addition, the rods and cones are not evenly distributed in the retina. For instance, there are more cones to perceive red light than there are to perceive other frequencies. In addition, the retinal region called the fovea corresponds to a special adaptation containing exclusively cones, which are densely packed and support the sharpest vision. The retinal fovea is essential for all activities that require the ability to perceive visual detail, such as reading and driving [KNFJ16]. It can be argued that this is the most important structure of the human eye that stereoscopic displays need to support when it comes to the perception of visual detail. Although each human eye has a field of view of nearly 180 degrees, the fovea covers only 15 degrees of the total field of view of the eye [NG16]. In the fovea there are 17,000 cones per square degree [KNFJ16], meaning a resolution of 130 x 130 pixels per degree would be necessary for a display to support the sharpest human vision, as captured by the fovea. For this reason, we propose to set the baseline for supporting high visual detail in the central region of the HVS at 130 pixels per degree. This may not seem like a particularly large number, but it is about twice as much as the overall proposed by some manufacturers [Wik15, Kan15]. For instance, AMD [Kan15] has produced an estimate of 116 million pixels (116 MPixels) to support stereoscopic vision. These 116 MPixels are distributed over the vertical and horizontal fields of view assuming a single display would be used to cover the complete field of view of a viewer (which is about 200 degrees horizontally and 135 degrees vertically), including the shared region between both eyes. Their estimate of display resolution to support overall stereoscopic vision is 60 ppd. However, in regards to the capabilities of the HVS, it could be argued that pixel density should be higher around the center of the field of view, where the visual field of both eyes overlap, and should be lower in the sides, in areas that support peripheral vision.

5.2 Field of View

The entire human binocular field of view (FOV) extends 190 degrees horizontally and 135 degrees vertically [HR95, SDM12], so we propose to use these figures as the baseline for supporting the complete horizontal and vertical binocular fields of view. In terms of coverage of the visual field of view, most display manufacturers have focused on two different approaches to support immersive AR and VR: either to cover a large part of the field of view at the expense of a lower resolution, or cover a limited part of the field of view and provide a good screen resolution. Because of the presence of the nasal block, which reduces the field of view in the region between the eyes, each individual eye's FOV is about 170 degrees horizontally [SDM12]. Stereoscopic displays do not need to cover the entire 170 degree field of view for each eye separately, because the eyes share a good portion of the entire field of view (114 out of 190 degrees) [HR95]. The eyes capture detail in a small region of the entire field of view. However, because the eyes move around to discover detail, for example, when we are reading, the region that the HVS covers to find sharp details includes eye individual eye rotations with a maximum amplitude of about 80 degrees, so the region that is suitable for viewing sharp details and stereo vision covers about 50 % of the whole field of view [Ful96]. This region occupies the center of the field of view. Therefore, two regions can be distinguished in the HVS; one that usually supports the central and highly detailed vision requiring 130 pixels per degree over about 50 % of the shared (central) FOV and another that supports peripheral vision with only 70 pixels per degree required over the remaining (temporal) part of the entire field of view.

5.3 Assessment of Immersive AR and VR Equipment

We divide our assessment in three main categories: Head Mounted Displays (HMDs) for immersive AR/VR, RGB Video Capture (mainly relevant for AR systems) and Depth-sensing cameras, which are useful for gesture capture and interpretation and for capture of the users' short- and medium-range surroundings.

5.3.1 Assessment of Head Mounted Displays

There is a flurry of development in hardware designed to support stereo vision in immersive Virtual Reality and Augmented Reality through Head-Mounted Displays [Ave16, Vir15, Vuz11, Vuz15, Sam15, Eps15]. The main configurations that have been recently proposed by HMD manufacturers to support stereoscopic vision in Augmented Reality and immersive VR are analyzed in this section. In Table 12, we characterize HMDs based on their support of a wide field of view and in terms of their pixel density, as a function of the capabilities of the human visual system. Display support of general vision, i.e., at a baseline of 60 ppds, varies from 17 % to 80 %. To support a baseline of 130 ppds for detailed vision, current display support is between 8 % and 17 % . The binocular horizontal FOV coverage ranges from 11 % for the most dense ppd displays to 49% for the widest FOV displays, this was estimated taking into account that the binocular field of view extends slightly over 190 degrees [HR95]. The vertical FOV coverage goes from 3 % in some displays to 74 %. These larger amounts (with respect to the horizontal FOV coverage), stems from the fact that the vertical FOV is 135 degrees, significantly smaller than the 190 degrees covered horizontally. On a separate issue, from the data shown in Table 12, it can be clearly concluded that displays that provide a wide field-of-view support exhibit the lowest density of pixels per degree. To characterize visual support provided by these very different technologies, we have introduced a measure called the Overall Visual Support (OVS). The OVS is calculated as the average of the four main metrics found earlier: general vision pixel density support, detailed vision pixel density support, and the binocular horizontal and vertical field of view coverage. Note that under this metric, a monocular display such as the Vuzix M100 or Google's Glass, would achieve a maximum of 50 % binocular FOV coverage, as the missing display would be accounted as 0 % coverage. With this metric, we are able to characterize the current state of the art in the display hardware, where a value of 100 % would imply a complete match to the characteristics of the HVS in binocular systems. This measure shows that current displays support from 20 % to 37 % of the HVS capabilities when the field of view and sensor density are taken into consideration.

5.3.2 Assessment of RGB Video Cameras

In addition to HMDs, Augmented Reality applications heavily rely on the use of RGB video cameras to capture elements from the real world that may need to be integrated with computer generated imagery [Int14, Int15, Ste15, Ash14]. The capacity of these cameras to convert visual detail would also have an impact on the user's ability to perceive visual and peripheral detail while wearing an HMD. Two main groups of cameras can be used to capture the environment: RGB cameras and Depth-sensing cameras. In some cases, these cameras are already integrated in one device, however, for the sake of a cleaner analysis we will consider them separately. First, we characterize the support RGB cameras provide to the users in terms of support for general vision pixel density, detailed vision pixel density and capture of a wide field of view. In Table 13, we characterize RGB cameras based on their support of stereoscopic capture, the amount of detail these cameras capture, and the field of view they cover. The values shown in the general vision sensor density are measured based on the cameras horizontal resolution and field of view against the baseline of 60 ppd. The highest resolution cameras provide reach up to 74 % support and down to 19 % support. For detailed vision sensor sensitivity,

Head Mounted Display	Latency (Hz)	PPD	Pixel Dens. Gen.	Pixel Dens. Detail	FOV Horizontal	FOV Vertical	Binocular FOV Coverage	Overall Visual Support
Oculus Rift DK2	60	10.2	17 %	8 %	49 %	74 %	62 %	37 %
Samsung Gear VR	50	20.1	33 %	15 %	34 %	53 %	43 %	34 %
Avegant Glyph	60/120	33.5	56%	26 %	20 %	18 %	19 %	30 %
Vuzix Wrap 1200 VR	60	27.9	47 %	21 %	16 %	13 %	14 %	24 %
Vuzix M100 Mono AR	60	31.1	52 %	24 %	3 %	3 %	3 %	20 %
Epson Moverio BT 200	60	47.9	80 %	37 %	11 %	8 %	9 %	34 %

Table 12: Assessment of Binocular Vision Support Provided by Head Mounted Displays

the baseline is 130 ppd and the resulting coverage goes from 34 % at the highest end to 9 % at the lowest end. Binocular field of view coverage for RGB cameras displays much less variation, with ranges from 23 % up to 50 % of the 190 degrees baseline. In terms of the vertical FOV coverage the range is similar, 29 % to 47 %; however, those devices which rank high in the horizontal FOV coverage are not the same as those which rank high in the vertical FOV coverage. Finally, results show some of the latest cameras support from 33 % to 45 % of the ideal overall visual support. It is not possible, however, to simply choose the camera that provides the best Overall Visual Support and attach it to an HMD. It is important to take into account the amount of visual detail that the HMD supports when compared to each single camera. For instance, a camera that captures 74 % of the visual detail may not be of much benefit to the user if the camera is connected to a display that has a general vision pixel density support of only 17 %, such as the Oculus DK2. A similar point can be made with respect to Field of View.

5.3.3 Assessment of Depth Sensing Cameras

Our analysis of depth-sensing cameras is made from the point of view of depth sensors which are mounted on the user and focuses on the support these provide for free motion and tracking of hands, referred to as wide gestural space, for short- and medium-range depth estimation, for the combined depth-sensing support and for the combined depth sensing and wide field of action support. Depth-sensors are used for a variety of applications, but in this case, we concentrate on the use of depth-sensors as wearable or at least first-person perspective devices to support tasks and actions that depend on the correct interpretation of depth data [Mot15, Int14, Int15, Ste15, Ash14]. When we refer to a wide field of action, we hypothesize about the support the body-worn devices provide for a user moving his or her arms and hands in a wide space, or whether the user is restricted to move within a certain region where the sensor is functional. To measure short range depth support we measure the coverage of the sensors within the first meter, which is about the longest distance that a person can reach with extended arms. The medium-range depth support is estimated as the amount of depth coverage the device supports from the first meter within the first 35 meters (as defined by [SJK$^+$07]), excluding the short-range coverage region and is relevant to the estimated depth and placement of objects that are within the interaction space of the user. The combined depth-sensing support is estimated as the average of the short and the medium-range depth-sensing support. Table 14 shows the results of our analysis for a wide array of depth-sensing cameras.

In terms of general vision pixel density support (measured horizontally), depth-sensing cameras vary from 7 % to 38 %. Detailed vision support is between 3 % up to 18 %. These numbers are clearly insufficient for high detail capture and are dramatically lower than the figures of RGB sensors previously discussed. The reason for this may be two-fold: first, estimating depth from stereo involves an additional extra processing step; and second, depth-sensing technologies are newer than other technologies. Nevertheless, depth-sensing cameras provide valuable information for stereoscopic manipulation, as well as for combining depth-cues from the synthetic world with depth-cues from the outside world, for both near and distant located objects. These cameras provide a wide range of angular field of action support, from 31 % to 84 %, with the latter being much more preferable. There are two categories of depth-sensing cameras, some sup-

RGB Camera	Latency (Hz)	Stereo Video Capture	PPD	Sensor Density General	Sensor Density for Detail	Binocular FOV Coverage	Visual Capture Support
Intel Real Sense R200	30-200	No	27.4	46 %	21 %	34 %	34 %
Intel Real Sense F200	30/60	No	27.4	46 %	21 %	34 %	34 %
ZED-15	30-120	Yes	23.0	38 %	18 %	45 %	37 %
ZED-30	15	Yes	20.0	33 %	15 %	45 %	35 %
Microsoft Kinect V1	30	No	11.2	19 %	9 %	31 %	22 %
Microsoft Kinect V2	30	No	27.4	46 %	21 %	41 %	37 %

Table 13: Assessment of RGB Video Capture for Use in Immersive Augmented Reality.

Depth Camera	Latency (Hz)	Range min	Range max	PPD	Pixel Dens. Gen.	Angular FOA & FOV Capture	Gesture & Short-Range Capture	Medium-Range Capture	Short & Medium Distance Support
Leap Motion	200 max	2 cm	60 cm	4.3	7 %	84 %	58 %	2 %	30 %
Intel Real Sense R200	30/60	20 cm	1.2 m	10.4	17 %	34 %	80 %	3 %	41 %
Intel Real Sense F200	30-120	51 cm	4 m	8.1	14 %	43 %	50 %	10 %	30 %
ZED-15	15	1.5 m	20m	23.0	38 %	45 %	0 %	53 %	26 %
ZED-30	30	1.5 m	20m	20.0	33 %	45 %	0 %	53 %	26 %
Microsoft Kinect V1	30	40 cm	4.5 m	5.6	9 %	31 %	80 %	12 %	46 %
Microsoft Kinect V2	30	50 cm	8 m	7.3	12 %	41 %	70 %	21 %	46 %
ZED-15 + R200 *	15	20 cm	20m	N/A	N/A	40 %	80 %	56 %	68 %
* Hypothetical mix									

Table 14: Assessment of Depth Sensing Cameras Support for Gestural and Medium Distance Capture.

port exclusively short-range capture suitable for arm and hand gesture interpretation, and some are exclusively suitable for medium-range depth-sensing purposes. Currently, a mixture of technologies would be suitable to cover a significant range of depth-sensing. For instance, combining an Intel RealSense R200 with a ZED -15 camera would hypothetically allow the capture of depths fields from 0.2 to 20 meters, i.e., 68 % of the support needed for short and medium distance capture.

5.4 Reduction of visual acuity with aging

As we have discussed in Section 3, visual acuity decreases with age. The practical consequence of this fact is that most existing devices might initially be suitable enough for older age individuals. As technological advances move towards the ideal display resolutions described previously, younger segments of the population would increasingly benefit from them.

6 Conclusions

In this paper, we present a new approach in which we suggest that the schemes for evaluating stereo algorithms should be designed based on the specific requirements of the target application and the characteristics of the human visual system. We then applied this concept to the particular application of AR in outdoor environments. We have chosen outdoor environments in this research since augmented reality systems and stereo vision algorithms deal with more challenges in these environments due to external factors that cannot be easily controlled, such as the effect of shadow and lighting. As a result, a practical analysis on the performance of the stereo algorithms, in terms of *accuracy* and *processing time* as perceived by the HVS, was presented. The results over the masked regions did not show any significant benefit to the evaluation of the areas near the depth discontinuities and occluded regions; however, as mentioned previously, this might be merely an indication of the performance of the algorithms we selected for evaluation and can only be better analyzed by evaluating more algorithms within our model. In either case, we hypothesize that, due to the importance of occlusion and areas near depth discontinuities to the HVS for the perception of depth in AR applications, it might be reasonable to focus more on the regions that contain depth edges and their surroundings when designing or employing a stereo matching technique for an AR application. Validation of this hypothesis is a topic we would like to further investigate in the future research. In this study, we presented an experiment to show the effect of post processing steps and refinement and its effect on the performance of the stereo algorithms. In addition, we presented an experiment to show the importance of the hardware used in AR systems, their resolution and robustness to noise and proved how important those factors are despite the accuracy of the results achieved by the software and the stereo algorithms in such applications. Further on, we would like to assess the benefits of our model for other AR applications, such as underwater environments and explore other factors which may also affect the evaluation process, such as the effect of contrast and brightness.

Our assessment of equipment characteristics evaluates hardware based on the amount of visual detail that can be captured by the HVS and the field of view of cameras and sensors used for immersive AR/VR systems. In our evaluation of HMDs, we confirmed our perception that devices that cover a wide FOV provide relatively low visual detail support in terms of pixels per degree. Alternative devices which provide high levels of visual detail are restricted to displaying much smaller regions of the binocular FOV. In our evaluation for RGB cameras, which are mainly meant to support interactions in immersive AR systems, we present that the amount of visual detail provided is higher than that found in HMDs, and the binocular FOV coverage is much more standard, between 31 % and 45 % of the ideal support needed for visual capture of the users' surroundings. Based on our study, depth-sensing cameras are used in two distinct type of applications. In the short range, they are used to capture and interpret users' gestures and commands, and the angular field of action and field of view capture is between 31 % to 84 % of the 190 degree of action. In the medium range, depth sensors are used to estimate depths from 1 to 20 meters, covering up to 56 % of the medium range of capture. Short- and medium-range depth sensors can be combined to provide support both gesture and short distance estimation, as well as depth capture of the users' environment.

7 Acknowledgments

This research has been funded by the Research and Development Corporation (RDC) of Newfoundland and Labrador.

References

[ABB$^+$01] Ronald Azuma, Yohan Baillot, Reinhold Behringer, Steven Feiner, Simon Julier, and Blair MacIntyre, *Recent Advances in Augmented Reality*, IEEE Computer Graphics and Applications **21** (2001), no. 6, 34–47, ISSN 0272-1716, DOI 10.1109/38.963459.

[Ash14] James Ashley, *The Imaginative Universal*, `http://www.imaginativeuniversal.com/blog/post/2014/03/05/`, 2014, Last visited July 18th, 2016.

[Ave16] Avegant, *Video headset - avegant glyph*, `https://www.avegant.com/product`, 2016, Last December 8th, 2016.

[Cor12] Tom Cornsweet, *Visual Perception*, Elsevier Science, Burlington, 2012, ISBN 978-0-323-14821-4.

[DM96] David Drasic and Paul Milgram, *Perceptual issues in augmented reality*, Stereoscopic Displays and Virtual Reality Systems III (Bellingham, Wash.) (Mark T. Bolas, Scott S. Fisher, and John O. Merritt, eds.), vol. 2653, SPIE, 1996, DOI 10.1117/12.237425, ISBN 0-8194-2027-1.

[DS14] Arindam Dey and Christian Sandor, *Lessons learned: Evaluating visualizations for occluded objects in handheld augmented reality*, International Journal of Human-Computer Studies **72** (2014), no. 10-11, 704–716, ISSN 1071-5819, DOI 10.1016/j.ijhcs.2014.04.001.

Citation

Bahar Pourazar and Oscar Meruvia-Pastor, *A Comprehensive Framework for Evaluation of Stereo Correspondence Solutions in Immersive Augmented and Virtual Realities*, Journal of Virtual Reality and Broadcasting, 13(2016), no. 2, December 2016, urn:nbn:de:0009-6-44179, DOI 10.20385/1860-2037/13.2016.2, ISSN 1860-2037.

[Eps15] Epson, *Developer's Guide to Moverio Smart Eyewear*, `http://www.epson.com/alf_upload/pdfs/brochure_moverio.pdf`, 2015, Last visited July 18th, 2016.

[FMHW97] Steven Feiner, Blair MacIntyre, Tobias Höllerer, and Anthony Webster, *A touring machine: Prototyping 3D mobile augmented reality systems for exploring the urban environment*, Personal Technologies **1** (1997), no. 4, 208–217, ISSN 1617-4909, DOI 10.1007/BF01682023.

[Ful96] James H. Fuller, *Eye position and target amplitude effects on human visual saccadic latencies*, Experimental Brain Research **109** (1996), no. 3, 457–466, ISSN 0014-4819, DOI 10.1007/BF00229630.

[GLU12] Andreas Geiger, Philip Lenz, and Raquel Urtasun, *Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, DOI 10.1109/CVPR.2012.6248074, pp. 3354–3361, ISBN 978-1-4673-1226-4.

[Goo13] Google Inc., *Google AR glasses*, `http://www.google.com/glass/start/`, 2013, Last visited July 18th, 2016.

[GS06] L. Garnham and J. J. Sloper, *Effect of age on adult stereoacuity as measured by different types of stereotest*, British Journal of Ophthalmology **90** (2006), no. 1, 91–95, ISSN 0007-1161, DOI 10.1136/bjo.2005.077719.

[Hec98] Eugene Hecht, *Optics*, 3rd ed., Addison-Wesley, Reading, Mass., 1998, ISBN 0-201-30425-2.

[Hir08] Heiko Hirschmüller, *Stereo Processing by Semiglobal Matching and Mutual Information*, IEEE Transactions on Pattern Analysis and Machine Intelligence **30** (2008), no. 2, 328–341, ISSN 0162-8828, DOI 10.1109/TPAMI.2007.1166.

[HP00] Aaron Hertzmann and Ken Perlin, *Painterly rendering for video and interaction*, NPAR '00 Proceedings of the 1st international symposium on Non-photorealistic animation and rendering (New York, NY), ACM, 2000, DOI 10.1145/340916.340917, pp. 7–12, ISBN 1-58113-277-8.

[HR95] Ian P. Howard and Brian J. Rogers, *Binocular vision and stereopsis*, Oxford psychology series, vol. 29, Oxford University Press, Oxford, 1995, ISBN 0-19-508476-4.

[Int14] Intel Corporation, *Intel realsense technology: Sdk guidelines*, version 2 ed., 2014, `https://software.intel.com/sites/default/files/managed/27/50/Intel%20RealSense%20SDK%20Design%20Guidelines%20F200%20v2.pdf`, Last visited July 18th, 2016.

[Int15] Intel Corporation, *Sdk design guidelines: Intel realsense camera (r200)*, version 1.1 ed., 2015, `https://software.intel.com/sites/default/files/Intel%20RealSense%20SDK%20Design%20Guidelines%20R200%20v1_1.pdf`, Last visited July 18th, 2016.

[JW05] Christian Jerome and Bob Witmer, *The Perception and Estimation of Egocentric Distance in Real and Augmented Reality Environments*, Proceedings of the Human Factors and Ergonomics Society Annual Meeting **49** (2005), no. 26, 2249–2252, ISSN 1541-9312, DOI 10.1177/154193120504902607.

[Kan15] David Kanter, *Graphics processing requirements for enabling immersive vr*, Tech. report, AMD, June 2015, `http://amd-dev.wpengine.netdna-cdn.com/wordpress/media/2012/10/gr_proc_req_for_enabling_immer_VR.pdf`.

[KNFJ16] Helga Kolb, Ralph Nelson, Eduardo Fernandez, and Bryan Jones, *Webvision: The organization of the retina and visual system*, `http://webvision.med.utah.edu/`, 2016, Last visited July 18th, 2016.

[KSF10] Ernst Kruijff, J. Edward Swan II, and Steven Feiner, *Perceptual issues in augmented reality revisited*, 9th IEEE International Symposium on Mixed and Augmented Reality (ISMAR), IEEE, 2010, DOI 10.1109/ISMAR.2010.5643530, pp. 3–12, ISBN 978-1-4244-9343-2.

[Liv05] Mark A. Livingston, *Evaluating human factors in augmented reality systems*, IEEE Computer Graphics and Applications **25** (2005), no. 6, 6–9, ISSN 0272-1716, DOI 10.1109/MCG.2005.130.

[Mot15] Leap Motion, *Leap motion for virtual reality beta*, `https://www.leapmotion.com/product/vr`, 2015, Last visited July 18th, 2016.

[MSZ$^+$11] Xing Mei, Xun Sun, Mingcai Zhou, Shaohui Jiao, Haitao Wang, and Xiaopeng Zhang, *On building an accurate stereo matching system on graphics hardware*, 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), IEEE, 2011, DOI 10.1109/ICCVW.2011.6130280, pp. 467–474, ISBN 978-1-4673-0062-9.

[NG16] C. R. Nave and Georgia State University, *Hyperphysics - the retina*, `http://hyperphysics.phy-astr.gsu.edu/hbase/vision/retina.html`, 2016, Last visited July 14th, 2016.

[Pfa00] Jonathan David Pfautz, *Depth Perception in Computer Graphics*, Ph.D. thesis, University of Cambridge, 2000.

[Rea83] R. W. Reading, *Binocular Vision: foundations and applications*, Butterworths, Boston, 1983, ISBN 0-409-95033-5.

[Sam15] Samsung, *Samsung gear vr*, http://www.samsung.com/global/galaxy/wearables/gear-vr/, 2015, Last visited July 18th, 2016.

[SB12] Daniel Scharstein and Anna Blasiak, *Middlebury stereo evaluation - version 2*, http://vision.middlebury.edu/stereo/eval/, July 2012, Last visited July 14th, 2016.

[SDM12] Peter J. Savino and Helen V. Danesh-Meyer (eds.), *Neuro-ophthalmology*, 2nd ed., Color atlas & synopsis of clinical ophthalmology, Wolters Kluwer/Lippincott Williams & Wilkin, Philadelphia, Pa., 2012, ISBN 1-60913-266-1.

[SJK$^+$07] J. Edward Swan II, Adam Jones, Eric Kolstad, Mark A. Livingston, and Harvey S. Smallman, *Egocentric Depth Judgments in Optical, See-Through Augmented Reality*, IEEE Transactions on Visualization and Computer Graphics **13** (2007), no. 3, 429–442, ISSN 1077-2626, DOI 10.1109/TVCG.2007.1035.

[SS02] Daniel Scharstein and Richard Szeliski, *A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms*, International Journal of Computer Vision **47** (2002), no. 1, 7–42, ISSN 0920-5691, DOI 10.1023/A:1014573219977.

[SSJE12] Gurjot Singh, J. Edward Swan II, J. Adam Jones, and Stephen R. Ellis, *Depth judgments by reaching and matching in near-field augmented reality*, IEEE Virtual Reality Workshops (VRW), IEEE, 2012, DOI 10.1109/VR.2012.6180933, pp. 165–166, ISBN 978-1-4673-1247-9.

[Ste15] StereoLabs, *ZED depth sensing and camera tracking*, www.stereolabs.com/\discretionary{-}{}{}zed/specs, 2015, Last visited July 18th, 2016.

[Sze11] Richard Szeliski, *Computer vision: algorithms and applications*, Springer, London, 2011, ISBN 978-1-84882-934-3.

[Vir15] Virtual Reality Times, *List of all virtual reality headsets under development*, http://www.virtualrealitytimes.com/2015/03/19/, 2015, Last visited July 18th, 2016.

[Vuz11] Vuzix, *Wrap Video Eyewear Comparison Chart*, https://www.inition.co.uk/wp-content/uploads/2014/03/Vuzix_CES_Brochure_20111.pdf#page=4, p. 4, 2011, Last Visited December 8th, 2016.

[Vuz15] Vuzix, *M100 smart glasses*, https://www.vuzix.com/Products/M100-Smart-Glasses, 2015, Last visited July 18th, 2016.

[Wik15] Wikipedia, *Retina display*, https://en.wikipedia.org/wiki/Retina_Display, Nov 2015, Last visited July 18th, 2016.

[WRMW95] John P. Wann, Simon Rushton, and Mark Mon-Williams, *Natural problems for stereoscopic depth perception in virtual environments*, Vision Research **35** (1995), no. 19, 2731–2736, ISSN 0042-6989, DOI 10.1016/0042-6989(95)00018-U.

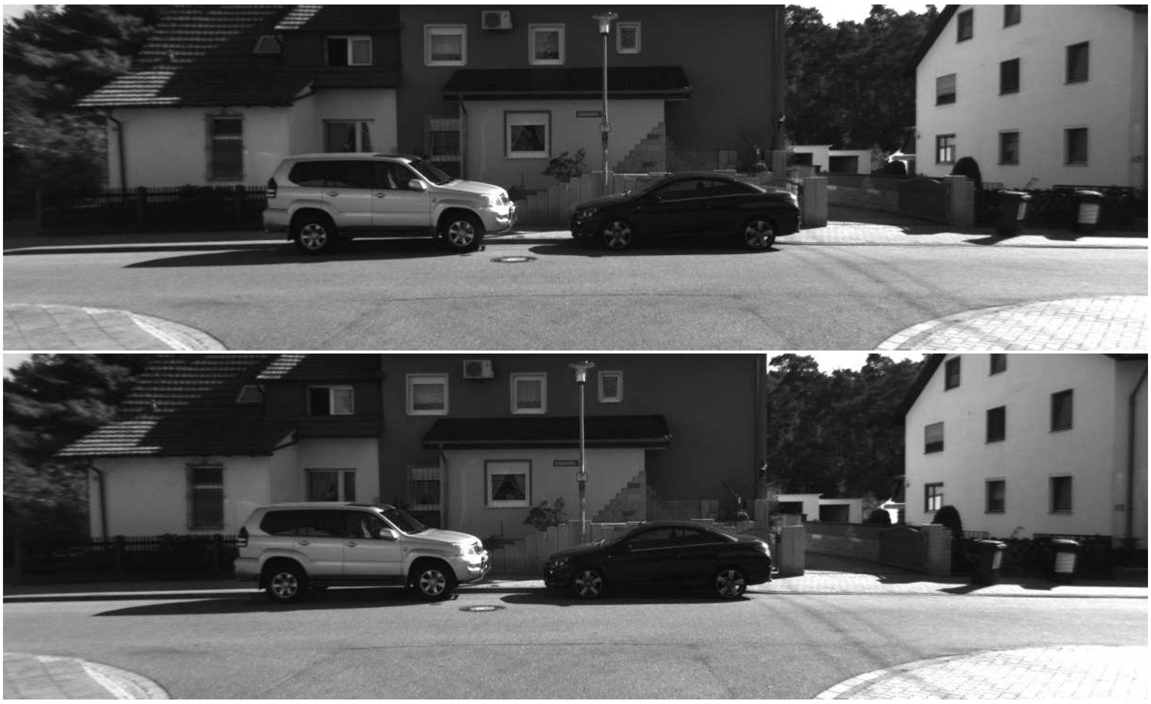


Figure 13: Sample outdoor scene from KITTI stereo dataset. Top: left image. Bottom: right image

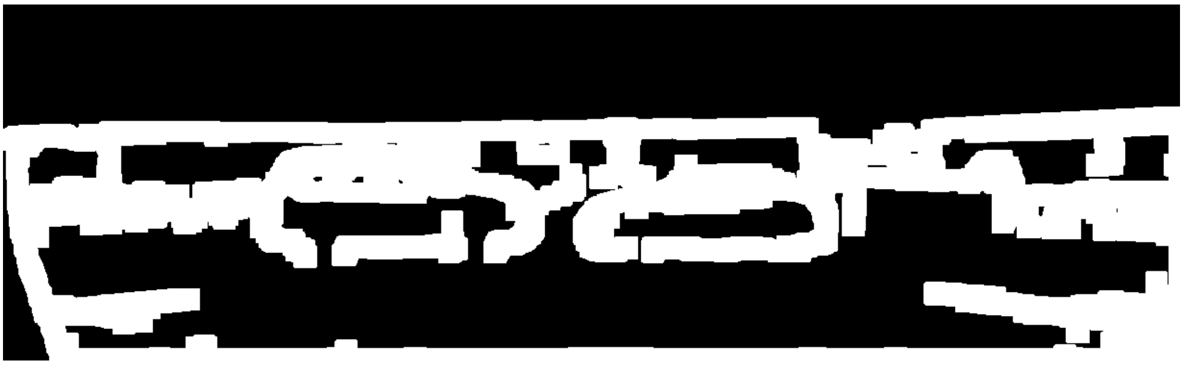


Figure 14: The mask of depth edges and their surrounding regions for Figure 13

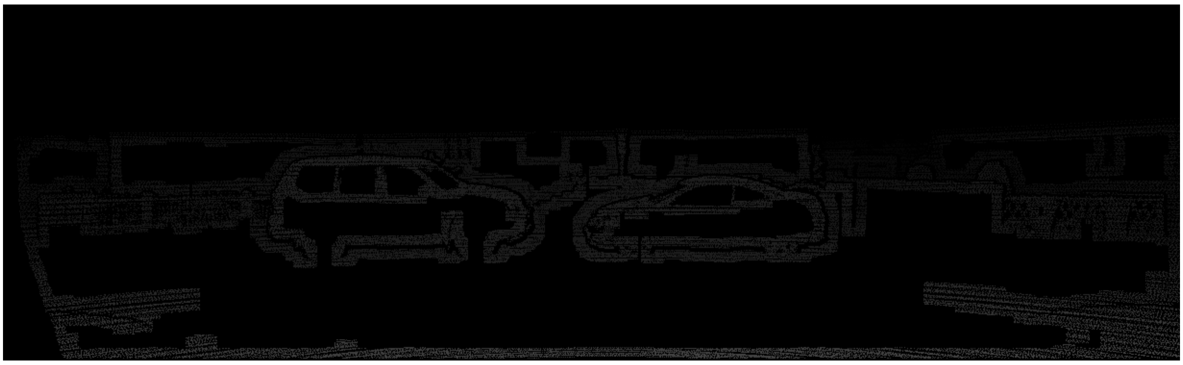


Figure 15: Masked ground truth for Figure 13

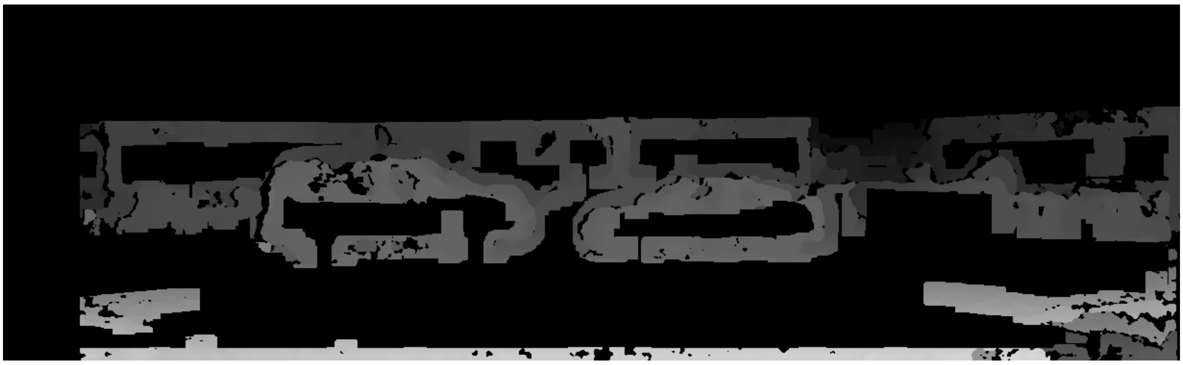


Figure 16: Masked disparity by SGBM for Figure 13

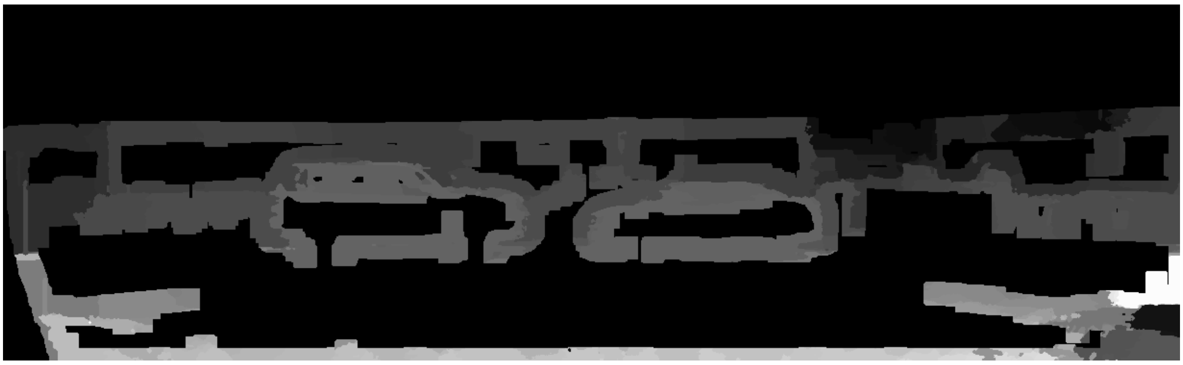


Figure 17: Masked disparity by ADCensusB for Figure 13