Studies for a Spatial See-Through Augmented Reality System to support Manufacturing Operations

Diogo Vicente Cleto dos Santos de Sousa Rodrigues

diogo.cleto@tecnico.ulisboa.pt

Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal
December 2021

ABSTRACT

See-through devices are a particular type of Augmented Reality technology, developed and used with the aim of enhancing visual perception of the physical world with additive light.
The present research work was motivated by the objective of establishing the proof-of-concept of a proposed framework for a Spatial See-Through AR system supported by a transparent screen. The screen is responsible for displaying, in the correct position, virtual information generated by the computer, intended to enhance the world behind itself. The virtual content has the particularity of adapting and interacting with the user and the reality that surrounds it.
Two specific approaches were presented to achieve the defined objectives through different strategies. Each strategy relied on Computer vision, Monocular Depth Perception and 3D Monoscopy, from which different methods were reviewed and specifically developed and adapted to the problem statement.
The synergy of the three methods in both approaches was responsible for fulfilling the objective of allowing users to employ their natural spatial perception while interacting with digital information, without additionally resorting to any other device. Hence, installation of the system was expected to be minimally disruptive to users' work processes and shop floor layouts.
The system demonstrated its reliability by accurately displaying virtual elements relative to the position of the real object behind the screen. Regarding the users' comfort and well-being, some symptoms were pointed as potential issues after prolonged use.

Keywords: Augmented Reality, Spatial Optical See-Through, Computer Vision, Monocular Depth Estimation, 3D Monoscopy, Manufacturing.

1 Introduction

Implementing AR technology in manufacturing applications is a strong and growing area of research. As a result, it is globally used to provide simulations, assist manufacturing processes, and contribute to reducing lead times, lowering costs and enhancing quality. Even so, AR does not appear to be ready for industrial deployment in some areas, as there are still some issues to be addressed.

The prominence of technological advances in the manufacturing industry does not diminish the importance of human workers in this field. In this regard, user acceptance is crucial, as it may be responsible for hampering the technology's efficiency. Consequently, for this work, a user-focused approach will be taken in the development of the augmented reality system. As part of this approach, adversities identified as responsible for limiting the application of AR technologies in shop-floor operations and processes, were addressed.

It was proposed a conceptual framework for an AR system capable of enabling the user to interact effectively with virtual information during manufacturing operations. Furthermore, it should achieve a sense of presence in the real world by allowing users to employ their natural spatial perception. Installation of the system is expected to be minimally disruptive to shop floor layouts and users' work processes.

Explicitly, the project aims to establish the concept's proof-of-concept by studying the development of an Spatial See-Through system grounded by methods from several subjects and supported by a transparent screen. These methods include: Computer vision, Monocular Depth Perception and 3D Monoscopy. The screen will be responsible for displaying, in the correct position, information, images or any other type of virtual scene generated by the computer, intended to enhance the world behind the screen. The virtual content will have the particularity of adapting and interacting with the reality that surrounds it. In other words, any user movements or changes in space must be interpreted and processed by the computer so that the image on the screen remains updated and adjusted to the panorama of each moment.

The remaining of this document is organized as follows. Chapter 2 develops and adapts methods from Computer Vision, Monocular Depth Estimation and 3D Monoscopy. Chapter 3 describes the implementation of two approaches for the project. Chapter 4 presents and discusses the results. Finally, Chapter 5 reviews the achievements of the research work and presents the future viable applications for the system.

2 Methods

2.1 Computer Vision

2.1.1 Image Subtraction

The first method presented portrays an approach that uses arithmetic applied to images. Typically, within the research works reviewed, two approaches are used to achieve this end. More specifically, background scenario subtraction and temporal differentiation. Respectively, the first approach involves subtracting a frame that portrays only the background from original frames that capture the presence of the object of interest. In the second approach, more suitable for applications where the background is not constant, the subtraction is done between consecutive frames throughout the captured video.

After The image subtraction process itself is ended, a perfect selection of the area corresponding to the object is rarely immediately obtained, without the final image being subjected to additional treatment beforehand.Some usually implemented measures stand out: Adapt the defined threshold value to impose the white color on the pixels of the final image; Impose the white color on all black areas that are completely surrounded by white pixels; Force the black color for all white regions that are connected to other regions of the same color, by pixels that do not establish contact in both vertical and horizontal directions simultaneously. This measurement will be applied only to white regions whose area is less than a certain number of pixels;Eliminate all independent white regions that have an area smaller than the largest white region in the image.

Regarding this method some factors that must be taken into account in order to achieve quality final results. Of these factors, the following stand out:

- The attention required to variations in luminosity in the space where the tests are being carried out;
- The care to be taken with shadows, since the presence of shadows in the image can be interpreted as the presence of an important object, or it can even deform the selection of the main object itself;
- Detection of transparent or overly reflective objects (mirrors, for example) may be difficult because they do not have a color that distinguishes them from the surrounding scenery;
- the method's performance depends on the quality of the background and on the camera's resolution.

In general, this method proves to be faithful in detecting the user for conditions where the background was fixed and was simple in terms of texture and color, contrasting with the person in front. On the other hand, under the same conditions using the second approach, where subtraction is done between consecutive frames along the captured video, an additional problem emerged. The fact that the subtraction is done between consecutive frames implies that there has to be constant movement in the captured video for an object to be identified. What was found, in the context of the tests carried out, was that when the user was still or his movement was not detected by the camera, the entire image was considered the background and therefore the person's head was no longer identified. This is understandable since during several frames in a row the image remained the same and therefore the subtraction will result in a null matrix during those moments. That said, this approach is excluded for future implementation in the project.

Nevertheless, the first approach to this method, in the modes in which it was presented, is quite unreliable in conditions of variable scenarios, complex in terms of textures and colors. It is for this reason also unsuitable for future implementation. Reinforcing this decision even further, it should be noted that it will be necessary to opt for a computer vision method that will subsequently allow the identification of the position of the user's eyes with some accuracy. By the image subtraction method, the position of the eyes would have to be estimated and would be subject to many assumptions and associated errors. In case the final image region includes more than from the neck up, this task would be even more complicated.

2.1.2 Face Tracking

For the second method in Computer Vision's area, the field of facial identification and tracking is used. The Viola-Jones method[1] is one of the most used developed (e.g. mobile phones applications). It is a robust and computationally undemanding facial detection algorithm that is still globally studied and implemented in highly successful software.

Although it is possible to apply the Viola-Jones method to each frame, it becomes very demanding at a computational level, especially considering that a camera records videos at dozens of frames per second, at least. It would be inconceivable for most applications and especially for those intended for systems that work in real-time.

Therefore, it was necessary to explore an algorithm developed to apply Viola-Jones only once. The Kanade Lucas Tomasi (KLT) algorithm was introduced by Lucas and Kanade lucas:kanade and later developed by Kanade and Tomasi tomasi:kanade. It is characterised by detecting distinct points of a certain object in an image, which are recognizable by their texture, to then be located along the frames track$_3$, *robust*2.

That said, the fact that, for various reasons, the feature points initially identified may not be located in consecutive frames is highlighted. These reasons, in most cases, are associated with the partial concealment of the face, or even cases where the system has not been able to predict the displacement of the respective point. Thus, the tracking algorithm will proceed in the same way, but with fewer points to control. Note that in the event that a point has been covered and reappears in the image, it won't be included again in the algorithm's detection cycle. Consequently, it is possible that the number of points at a certain point is close to zero or even zero, which implies that the face is no longer located and detected.

The proposed method was tested. Facial detection was only performed once, ideally when the head was facing the webcam. Taking this into account, during the test time it was possible to rotate the head freely without the detection failing. This is because the number of identified feature points was large enough. Thus, it is possible to continue to predict the location of the user's face even when some of the points are no longer controlled by the algorithm, due to head movements. However, detection ends up failing when too many movements are performed in a single test, causing a gradual but significant reduction in the number of points under control, implying a break in the tracking sequence.

The speed of video presentation with the face and characteristic points identified is stable and perfectly compatible with real-time applications, as the delay in reaction and processing time is practically imperceptible.

The surrounding environment was not presented as an obstacle to the success of the program, so the tests did not highlight any aspect that could interfere with the performance of the algorithms.

There was, however, one problem that stood out during the tests. As the number of feature points continuously decreases or even gets to zero, implies a complete collapse in the face tracking process.

Therefore, there is a need to create a cycle that allows the area of interest to be located with some regularity. The tests showed that there is only a risk of collapse when the number of points is very low, meaning that the need is not periodic over time, but in terms of the number of feature points that are still being controlled by the algorithm. Thus, the cycle will be defined by re-detecting the user's face whenever a minimum of feature points is not met, so that they are renewed and inserted in the sequence of tracking more and new points.

The feasibility of implementing the Viola-Jones and KLT algorithms together as a way to respond to the first need to detect the user's faces is evident.

2.2 Depth Estimation

2.2.1 Exponential Fit

Using the results of a literature review, several strategies are identified that may incorporate the future method that will be used in this study. The method to be developed will be quite demanding in terms of requirements. The main reason for this is that the future SST system will integrate different methods from the three areas covered in this chapter. The Computer Vision method, by itself, comprises the application of multiple algorithms and with the additional and simultaneous operation of the Depth Perception and Screen Output methods, the more demanding the system will become at a computational level. Therefore, it is necessary to guarantee that the integral system will be prepared to work in real-time with high precision results. Being aware that the properties of the computer's graphic capacity to be used when testing the final system are still unknown, it is necessary to define, from now on, the respective requirements of the method to be developed in this area. Although they won't be able to certify the validity of the method, these tests will be critical to allowing at least a proof-of-concept to be established. Criteria will be established based on the needs of the project and on the research findings.

- For the first criterion, it is defined that the method should be based on a geometric and mathematical relationship involving a maximum of three parameters.
- The parameters must be values obtained directly from the camera's characteristics, or any other object/equipment involved in the process. They may also be extracted directly from the captured images.
- The method should have a minimum precision of 96% in the final results obtained.

It was decided to establish that this method should result in a mathematical relationship involving the longitudinal distance of the object to the camera (cm), the physical dimension of the object (cm) and the dimension of the object in the image plane (pixels). This will ensure that, regardless of the object and its dimension, only one relation is suitable to assess the depth, as long as it is fully captured by the camera.

The expression relating the parameters is therefore a function in $\mathbb{R}^3$. However, it is possible to simplify the function and convert it to $\mathbb{R}^2$ by creating a variable that is defined by the ratio between the pixel length of the object in the image and the

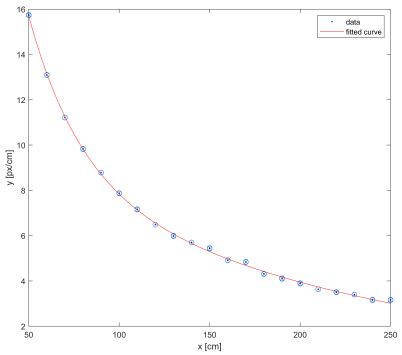

Figure 1. Compound Exponential Model fitted to the Data set.

Figure 2. First Laptop successfully adapted into a transparent monitor .

physical length in centimeters. This variable is characterised by corresponding to the inverse of the reality representation factor (δ_x) in the image plane, in cm/px, in the object plane, parallel to the image plane. Thus obtaining a relationship between only two parameters, since one of them is dependent on the aforementioned lengths and the other is the distance from the object to the camera.

Two objects, of different heights, were photographed at twenty-one different longitudinal distances from the camera each. The images were later analyzed individually and the height of each object, in pixels, was measured and associated with its respective longitudinal distance. For each measured height, the quotient between the dimension, in pixels, and the height of the respective object, in centimeters, was calculated. Thus, a set of forty-two ordered pairs (x,y) was obtained, where x refers to the distance from the object to the camera and y refers to the quotient defined for the respective distance x. Then, the ordered pairs were represented in an orthonormalized referential.

Regarding the data from the test carried out, the following aspects are highlighted:

- The longitudinal distances considered during the test were set between the lengths 50 and 250 centimeters, inclusive, with increments of 10 centimeters separating each one;
- The physical heights of the objects are respectively 21.75 and 16.50 centimeters;
- The resolution of captured photos is 1280 x 720 and the camera has a resolution of 0.9 MP.

A non-linear regression was carried out, on the point cloud representation of the data obtained from the test. With all criteria considered, an exponential regression was chosen as it meets the requirements presented (Figure 1).

$$f(x) = a \cdot e^{b \cdot x} + c \cdot e^{d \cdot x} \tag{1}$$

where a, b, c and d are the independent coefficients to be determined.

The following results were obtained:

Coefficients	(with 95% confidence intervals)
$a = 41.12$	$(37.34; 44.90)$
$b = -0.0351$	$(-0.03787; -0.03233)$
$c = 11.17$	$(10.39; 11.94)$
$d = -0.005246$	$(-0.005583; -0.004908)$

The coefficients of determination, R^2 and Adjusted R^2, were calculated:

$$R^2 = 99.9452\% \quad ; \quad Adjusted\ R^2 = 99.9392\%$$

As well as the Residual Standard Error and the Root Mean Square Error:

$$S = 0.0817\ px/cm \quad ; \quad RMSE = 0.0797\ px/cm$$

The presented results demonstrate that it is possible to implement a method that fulfills all previously established criteria.

2.3 Screen Output

After interpreting and analyzing all the data extracted from the images captured by the webcam, it becomes possible to proceed with methods that adapt the image on the transparent screen according to the user's field of view.

2.3.1 Line of Sight

The Line of Sight parametric equations will be suitable for whatever the position of the user or the object. For any movement that is detected on either side of the screen, the previous methods will be responsible for identifying and locating the new position of the respective object or of the user itself. Thus, the new coordinates will be extracted and updated in the parametric equations so that the line of sight is renewed over time.

Note that the equations fit into a three-dimensional coordinate system, where both the user and the object are defined by two points that establish the extremes of the line of sight, represented by a straight line. In the same reference, it is also possible to define the equation of the plane that represents the screen in its position and orientation in relation to the user and other objects.

The purpose of this method will be to dynamically display virtual markers on the transparent screen so that, from the user's point of view, they coincide with the objects of interest behind the screen. To successfully achieve this result, it is necessary to determine the intersection point of the line of sight with the transparent screen. The point $Q\ (Q_x, Q_y, Q_z)$ defines the position of the user that allows evaluating his/her point of view. The point $P\ (P_x, P_y, P_z)$ defines the position of a specific object on the other side of the screen.

he point of intersection between the plane α and the line h is defined by $I(I_x, I_y, I_z)$. The coordinates of point I are obtained by solving the following matrix equation:

$$\begin{pmatrix} n_y & -n_x & 0 \\ 0 & -n_z & n_y \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} I_x \\ I_y \\ I_z \end{pmatrix} = \begin{pmatrix} Q_x \cdot n_y - Q_y \cdot n_x \\ Q_z \cdot n_y - Q_y \cdot n_z \\ -d \end{pmatrix} \tag{2}$$

Where,

$$\begin{cases} n_x = P_x - Q_x \\ n_y = P_y - Q_y \\ n_z = P_z - Q_z \end{cases} \tag{3}$$

2.3.2 Virtual Real World

The method development in this section dedicated to Screen Output will be guided by the toe-in approach applied to the theory of monoscopy. However, toe-in is a process, considered by Oliver Kreylos[2] that gives incorrect results and causes discomfort to the user when compared to other methods.However, this is still a widely used method, and it appears as an adequate solution when the rendering engine or library does not provide off-axis rendering capabilities[3]. For this reason it is considered a versatile and suitable option for future implementation in the SST system.

To start it is necessary to determine the angle of view, α, for the position of the point of view (x, y). The angle of view, both vertical and horizontal, will be fundamental to mathematically define the 'frustum'.

$$cos(\alpha) = \frac{-x_R + x_R^2 + y_R^2}{\sqrt{x_R^2 - 2x_R^3 + x_R^4 + y_R^2 x_R^2 + y_R^2 - 2x_R y_R^2 + x_R^2 y_R^2 + y_R^4}} \tag{4}$$

where,

$$x_R = \frac{x + \frac{L}{2}}{L} = \frac{x}{L} + \frac{1}{2} \quad , \quad y_R = \frac{y + \frac{L}{2}}{L} = \frac{y}{L} + \frac{1}{2} \tag{5}$$

Where L is the screen's Width or Height, for the Top-view or Side-view, respectively.

Then it is necessary to determine the 'toe-in' angle that will be represented by the camera's rotation angle γ. This will also take into consideration the simplification previously implemented.

The angle γ is defined by the following relationship between the angle of view, α, and the angle θ:

$$\gamma = \frac{\alpha}{2} - \theta \tag{6}$$

where θ is the angle defined between vectors $\vec{v}$ and $\vec{UU'}$.

vector $\vec{UU'}$ originates from the point of view U and the norm and orientation of the segment joining point U to U'.

3 Implementation

Currently, a transparent OLED monitor is expensive and the costs associated with the acquisition of this type of device are not justified for the proof of concept behind this project. Thus, the alternative of using a conventional laptop and adapting it to the project context emerged. Descriptively, laptops donated to the project were used to establish the particularity of having a transparent monitor. With this solution, the developed programs would be executed directly on the laptop and projected on the respective monitor. Image capture would be performed through an external or internal webcam depending on the characteristics of the laptop used.

To implement the solution idealized in the previous paragraph, we proceeded with the task of disassembling the monitor from the laptop so that it was possible to remove the plastic cover and frame that surrounds the LCD. The first complete success of this operation was achieved on a computer without an integrated webcam. Transparency was achieved simultaneously with the usual functioning of the computer, however some disadvantages were highlighted:

- Without the support or frame that protected the LCD, the laptop was exposed to the outside environment and was weakened by any rotational movements of the screen.
- The laptop was left with the electrical wires too exposed, which could present some danger during its use.
- The view through the monitor was partially obstructed by the flat cable connecting to the LCD command printed circuit (CCFL).
- The polarizing filters present were responsible for not allowing the image of the environment behind the monitor to be completely clear.

In order to assess the reliability of the approaches that were developed, a transparent monitor was required to project the virtual content generated by the developed programs. Another requirement was that a camera should be integrated into the transparent monitor plane. Lastly, a computer should be set up to communicate constantly and simultaneously with the camera and screen.

The same operation could still be successfully reproduced on another laptop (Figure 2). For this, the results were different mainly due to the characteristics of the laptop itself. The list of disadvantages presented above has been shortened. Namely:

- The monitor incorporated a metal frame that covered a narrow frame of the monitor area.
- The electrical cables were not exposed and the rotation movements of the screen itself did not compromise the laptop's resistance.
- The flat cable connecting the LCD command printed circuit was not installed across the screen.
- The polarizing filters on this laptop had the same impact on image clearness as was visible through the monitor.
- The laptop integrated an internal webcam that remained functional after the monitor's transformation processes.

The characteristics of the second laptop described were identical to those of the computer used during the tests, which resulted in the Exponential Fit method, developed in Chapter 3, in the Depth Estimation section.

Regarding the features of the screen:

- Projection area dimensions: 34.4 X 19.2 [cm]
- It was ensured that the monitor plane was oriented perpendicular to the keyboard plane and, consequently, to the plane where the laptop was settled.

Regarding webcam features:

- Resolution: 1280 X 720 [px] and 0.9 MP
- The webcam is located centered with the screen's vertical axis of symmetry, 0.8 cm away from the top edge
- Horizontal capture angle: 79.42 (obtained empirically)
- Vertical capture angle: 50.27 (obtained empirically)

3.1 Line of Sight-based Approach

The first approach implemented consists of a synergistic combination of three methods, respective to each of the three areas presented in the previous chapter. For the Computer Vision area, Face Tracking was chosen, namely the Viola Jones, AdaBoost and KLT algorithms. These were applied sequentially and in a structured way, exactly as described in the respective method section. The Exponential Fit method developed and presented in the section dedicated to Depth Estimation was immediately recognized as the preferred method to integrate the system due to its simplicity and effectiveness. Finally, for this approach, the Screen Output area method to be implemented will be the Line of Sight Intersection, as the title of this section suggested.

The methods identified in the previous paragraph were implemented with the support of *MATLAB*. This software is not the most suitable for computer vision applications, compared to programming languages like C and derivatives. This fact is logical because *MATLAB* is an interpreter, therefore it has slower latency and processing times. Executions in C and derivatives are absolutely faster, or even eventually in Python. However, any of them implies a more time-consuming and demanding implementation than *MATLAB*[4]. So the interpreter was chosen, as it allows to conduct the proof of concept successfully from algorithms more elementary, intrinsic to the software[5]. *Simulink* also presents itself as a useful tool, however, it was not considered in this project for presenting less satisfactory results in research works with face tracking applications, similar to what is required in this project[4].

3.2 3D Monoscopy-based Approach

The second approach also involves three methods, respectively from each area of the previous section. The big distinction between the first and second approaches developed is the method chosen for Screen Output. Both use Viola Jones, AdaBoost and KLT algorithms, for the area of Computer Vision, and also use the Exponential Fit method, to estimate the depth of objects captured by the camera. This approach implements the Virtual Real World method.

Although the first two methods are common to the first approach, the modes in which they were implemented for the 3D Monoscopy-based approach are different. What distinguishes the way these methods were implemented differently is mainly the support software used to render the three-dimensional modeled environment on the transparent screen. We chose to use Unity 3D, due to its success in other research works reviewed dedicated to the exploration of augmented reality technologies[6,7] and the number of libraries available that aim to support computer vision and augmented reality applications. Unity 3D's user-friendly interface and tools are additionally an advantage that was highlighted which would make the implementation more straightforward. The use of Unity 3D implies that the code is written in C#, also allowing the latency and processing times to be faster, as already mentioned. Predicting that the real-time operation of the system will be successful.

Within this approach, an attempt was made to implement the three methods on a Server/Client configuration established between *MATLAB* and Unity 3D. This attempt was aimed at making constant communication between the two software. *MATLAB* was responsible for detecting and tracking the user's face, as well as estimating the coordinates of the point of view, relying on the work developed during the Line of Sight-based approach. The parameters to be updated in Unity 3D's virtual camera properties would be determined in *MATLAB* too. *MATLAB* was defined as client, and Unity 3D as server, to achieve this goal.

The attempt was eventually discarded because it was a demanding process at a computational level. This fact was evidenced by the inability to update more than one parameter in the camera properties. This communication solution would not be viable in a system operating in real time due to the response delay.

4 Results

In the 3D-Monoscopy approach the coordinates of each pixel displayed on the screen are not determined to exactly match the respective element of the surrounding environment, from the user's perspective. Thus, for user positions not exactly centered on the screen, regardless of its longitudinal distance, there are deviations between the position of the real object and its three-dimensional model on the screen (Figure 3).

A theoretical evaluation of the error between the displayed position of virtual objects on the screen and the real and correct position obtained by the Line of Sight method was carried out. The influence of the user's lateral and vertical displacements in relation to the camera was individually analyzed. The results are shown in Figure 4.

Quantitatively interpreting the absolute errors, it appears that the maximum error, for the area of analysis, corresponds to just over 2.5 centimeters. Considering that the screen used has a width of 34.4 centimeters, it implies a maximum relative error of 7.27% in relation to the dimensions of the displayed image. This error is obtained for a longitudinal distance of 50.00 centimeters and a horizontal displacement of 41.53 centimeters.

Since there is no standard threshold value that defines the acceptable error for the deviation of the images presented, a study was carried out to assess the space available for the user to move so that the visual deviation is never greater than 1%, 0.5% or 0.25% of his/her longitudinal distance to the camera. The results obtained are shown in Figure 5.

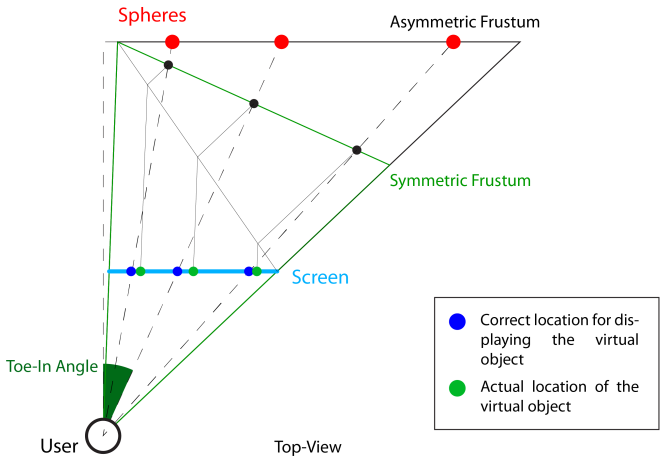

Figure 3. Comparison between the Line of Sight Intersection Method and the Toe-in Method.

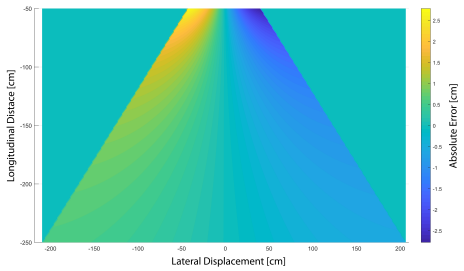

Figure 4. Absolute error of the virtual object position as a function of longitudinal distance and lateral displacement.

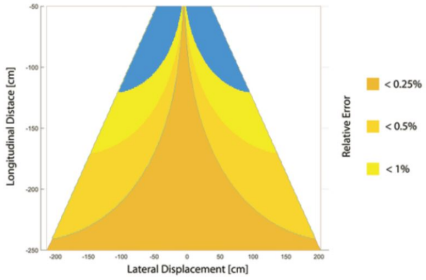

Figure 5. Region of space where visual deviation is less than 1%, 0.5% or 0.25%.

The results presented so far have been determined by always assuming the same fixed position of a particular object of interest. The absolute error was verified for objects that were at the longitudinal distance of the screen of 0.5, 1 and 2.5 meters. The results show that only the user's position has an influence on that deviation. This is a predictable conclusion as the shift is caused by the user's visual distortion of the screen due to their perception of perspective, of their current position.

It should be noted that the absolute error is not the same for all pixels in the image. It is possible to understand that in the vertical margins of the image, the absolute horizontal error is null and in the horizontal margins of the image, the absolute vertical error is null. It evolves to the maximum error value as it approaches the geometric center of the image.

In the same way that the presented position of the virtual objects presented deviations in relation to the correct position from the user's perspective, discrepancies in the dimension of the objects were also identified. By implementing the Exponential Fit method, it becomes possible to quantify the apparent dimensions of both, regardless of the longitudinal distance that separates the screen from the objects, and the objects from the user. The results obtained are shown in Figure 6. The horizontal axis represents the longitudinal distance from the user to the screen, in meters, and the vertical axis, the longitudinal distance from the object to the screen, in meters. Absolute errors are shown in centimeters.

Figure 6 presents theoretical results specific to an object used in the laboratory. Namely, was assumed a sphere with a diameter of 6.14 centimeters.

Counter-intuitively, the results conclude that for the study region, greater distances from the user and object to the camera contribute mainly to an increase in absolute errors. With the exception of the region defined by distance variations between 0.5 and 1.25 meters, approximately, where the referred monotony does not occur.

However, the errors that appear are less than 1.3 milimeters, which corresponds to less than 0.05% of the distance from the user and the object to the screen. Therefore, there is a need to assess the influence of object size on errors. In this way, exactly the same study was carried out for spheres with diameters of 10, 25 and 50 centimeters.The results establish a pattern between

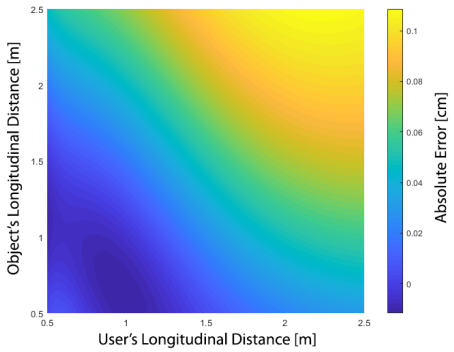

Figure 6. Absolute error for the apparent dimensions of the real objects compared their respective virtual representation.

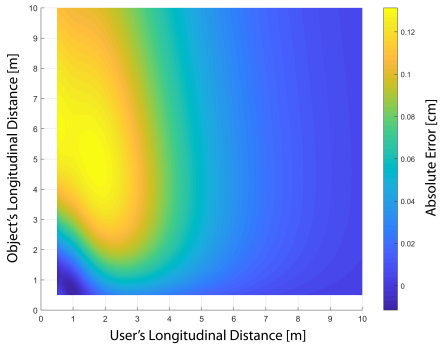

Figure 7. Absolute error for the apparent dimensions of the real objects compared their respective virtual representation until 10 meters of longitudinal distance.

the absolute errors of the apparent dimensions for different diameters of the sphere under study. The error range is the same for any diameter, by a difference of a scale factor. If the maximum error value determined for each of the diameters is considered, it is possible to conclude that the maximum is constant and corresponds to 1.6% of the diameter of the sphere under analysis. Considering that this maximum value is only verified when the user is at a distance of 2.5 meters from the screen, it implies that in the case of the larger diameter sphere, the absolute error corresponds to 0.32% of the user's longitudinal distance.

Following the study carried out for image deviation and since again there is no established threshold value that defines the acceptable error for this context, it will also be opportune to evaluate the diameters of the spheres so that the maximum absolute errors are less than the relative error levels as compared to the user's longitudinal distance.

Therefore, the error in the apparent dimension only exceeds 1%, 0.5% or 0.25% of the longitudinal distance, for spheres with diameters 156,250 cm, 78,125 cm and 39,063 cm, respectively.

The error in the initial sphere was also analyzed for longitudinal distances between 0.5 and 10 meters with the aim of verifying if the absolute error maintained its tendency to increase with increasing distances (Figure 7).

It can be seen, therefore, that the error variation reverses the behavior shown above, following a more intuitive evolution, namely, decreasing the error with the increase of longitudinal distances to the screen, similarly to the study on visual deviations.

The apparent perception of objects as a function of their distance from the user is estimated by a non-linear regression, more specifically, by the composition of two exponentials. The absolute error, in turn, is obtained by the difference between two double exponential regressions, which implies the involvement of four distinct exponential functions, which justifies the fact that there is no constant monotony in the results obtained. However, it is possible to state that for the area between 0 and 2 meters of the longitudinal distance of the user and 0 and 7 meters of the longitudinal distance of the object, the error has a mostly positive variation with the increase of either of the two distances. However, outside this region, that is, for values greater than 2 meters for the user and 7 meters for the objects, a tendency is expected to decrease the absolute error in the apparent dimensions.

5 Conclusion

In the area of Depth Estimation, a method specifically adapted to the project context was developed. By the results, the Exponential Fit method proved to be a viable solution with a high level of effectiveness, superior to the revised methods for similar purposes. The implementation of the method together with the application of Computer Vision algorithms contributed to make the user's interaction with the system possible with real-time response.

From the results discussed in Chapter 4, the reliability of the system was shown, but its direct use still has some issues associated with users' comfort and well-being after prolonged use. Symptoms such as headaches, eye fatigue, nausea and even epileptic seizures stand out. Although they have not been experienced, factors were identified that could contribute to their emergence.

This aspect suggests that the proposed framework for an AR system maintains its validity, however it requires the implementation of more refined systems or methods and solutions aimed at correcting and counteracting the factors that contribute to the listed symptoms. Possible examples that could be effective in combating the problems highlighted are based on technology developed by the automotive industry that integrates holographic augmented reality technology on the car's windshield.

Based on the results obtained, it is possible to predict the applicability of this type of technology. The main objective with the implementation of this solution is to allow the user to observe virtual information *in loco*. The vision of Industry 4.0 is coming through over the years so that virtual data is more constant in the industry and AR can enable it to be observed like no other technology. This system is a viable solution to do it with the least impact on the operation and the work environment, allowing it to be possible to produce and visualize a mental model of the user's work panorama, promoting easy data interpretation and decision making capabilities.

Explicitly, for the Manufacturing industry, the following applications are considered that may be viable:

- Monitoring of machines and production lines in real time from a transparent screen through which it is possible to observe the shop-floor. In this way, it is expected to be possible to control the production-rates of the machines and the production times, for example. It will also be a suitable solution to directly identify problems that may arise during production line processes and find out if there is already an operator on site to solve them.
- The previous example is still possible to expand for warehouse management activities where the technology appears to be promising. Namely, for order allocation, inventory management and order picking
- Support in the Machine-Tools area by installing a transparent screen in place of the machine window. By doing so, it may be possible to display real-time information directly superimposed on the operation, such as operating conditions or even virtual representations of the cutting tool movement. This solution may also be useful to guide the operator when the window is obstructed by the lubricant.
- Logistics management is anticipated to be an area of interest for SST system applications as it may allow faster decision-making and reduced error rates by simplifying the execution of natural logistics elements.

Additionally, it is foreseen the applicability of this project in areas that transcend manufacturing. Applications associated with exhibitions and museums, product displays and showcases, activities associated with tourism, culture and entertainment stand out, among which informative and didactic panels are particularly noteworthy. Sport will also be a possible application area, for information purposes in real time, during the course of events, which aim to monitor speeds and distances, among other information, live on the scene. This proposal can be easily extrapolated to any of the areas mentioned above. Applications of the SST system in the area of archeology may also be interesting for the virtual reconstruction of monuments, landscapes or scenery. Through the transparent screen, it is possible to visualize elements from certain eras superimposed on a modern environment, such as the reconstruction of ruins.

5.1 Final Remarks

The research work was carefully developed and structured with the aim of enhancing new projects. The exposed content aims to contextualize and provide the necessary bases for understanding the theme and the areas and scientific methods explored. Thus, it is expected that the project presented will enable the development of many applications in different areas of research.

References

1. .

2. *Kreylos, O. Good stereo vs. bad stereo.*

3. *Meindl, L.* Omnidirectional stereo rendering of virtual environments. *Ph.D. thesis, Wien (2015).*

4. *Luijten, H. Basics of color based computer vision implemented in matlab.* Tech. Eindhoven, Dep. Mech. Eng. Dyn. Control. Technol. Group, Eindh. *1–24 (2005).*

5. *Goyal, K., Agarwal, K. @AND@ Kumar, R. Face detection and tracking: Using opencv. In* 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, vol. 1, 474–478 (IEEE, 2017).*

6. *Kompaniets, A., Chemerys, H. @AND@ Krasheninnik, I. Using 3d modelling in design training simulator with augmented reality. (2020).*

7. *Smith, M., Maiti, A., Maxwell, A. D. @AND@ Kist, A. A. Using unity 3d as the augmented reality framework for remote access laboratories. In* International Conference on Remote Engineering and Virtual Instrumentation*, 581–590 (Springer, 2018).*