technologies

MDPI

Article

Perspective Preserving Solution for Quasi-Orthoscopic Video See-Through HMDs

Fabrizio Cutolo *, Umberto Fontana and Vincenzo Ferrari

Information Engineering Department, University of Pisa, 56122 Pisa, Italy; umbertofontana93@gmail.com (U.F.); vincenzo.ferrari@unipi.it (V.F.)

* Correspondence: fabrizio.cutolo@endocas.unipi.it; Tel.: +39-050-995-689

Received: 18 December 2017; Accepted: 10 January 2018; Published: 13 January 2018

Abstract: In non-orthoscopic video see-through (VST) head-mounted displays (HMDs), depth perception through stereopsis is adversely affected by sources of spatial perception errors. Solutions for parallax-free and orthoscopic VST HMDs were considered to ensure proper space perception but at expenses of an increased bulkiness and weight. In this work, we present a hybrid video-optical see-through HMD the geometry of which explicitly violates the rigorous conditions of orthostereoscopy. For properly recovering natural stereo fusion of the scene within the personal space in a region around a predefined distance from the observer, we partially resolve the eye-camera parallax by warping the camera images through a perspective preserving homography that accounts for the geometry of the VST HMD and refers to such distance. For validating our solution; we conducted objective and subjective tests. The goal of the tests was to assess the efficacy of our solution in recovering natural depth perception in the space around said reference distance. The results obtained showed that the quasi-orthoscopic setting of the HMD; together with the perspective preserving image warping; allow the recovering of a correct perception of the relative depths. The perceived distortion of space around the reference plane proved to be not as severe as predicted by the mathematical models.

Keywords: video see-through head-mounted displays; orthoscopy; perspective-preserving homography; stereo fusion

1. Introduction

Augmented reality (AR) systems based on head-mounted displays (HMDs) intrinsically provide the user with an egocentric viewpoint and represent the most ergonomic and efficient solution for aiding manual tasks performed under direct vision [1]. AR HMDs are commonly classified according to the AR paradigm they implement: video see-through (VST) or optical see-through (OST). In binocular VST HMDs, the view of the real world is captured by a pair of stereo cameras rigidly anchored to the visor with an anthropometric interaxial distance. The stereo views of the world are presented onto the HMD after being coherently combined with the virtual content [2].

By contrast, in OST HMDs, the user's direct view of the world is preserved. The fundamental OST paradigm in HMDs is still the same as that described by Benton (e.g., Google Glass, Microsoft HoloLens, Epson Moverio, Lumus Optical) [3]. The user's own view of the real world is herein augmented by projecting the virtual information on a beam combiner and then into the user's line of sight [4].

Although the OST HMDs were once at the leading edge of the AR research, their degree of adoption and diffusion slowed down over the years due to technological and human-factor limitations. Just to mention a few of them: the presence of a small augmentable field of view, the reduced brightness offered by standard LCOS micro displays, the perceptual conflicts between the 3D real world and the 2D virtual image and the need for accurate and robust eye-to-display calibrations [5].

Some of the technological limitations, like the small field of view, are being and will be likely overcome along with the technological progress. The remaining limitations are harder to cope with.

The pixel-wise video mixing technology that underpins the VST paradigm can offer high geometric coherence between virtual and real content. The main reasons for it are: unlike OST displays, the absence of a user-specific eye-to-display calibration routine; the possibility of rendering synchronously real scene and the virtual content, whereas in OST displays there is an intrinsic lag between the immediate perception of the real scene and the appearance of the virtual elements. From a perceptual standpoint, in VST systems the visual experience of both the real and virtual content can be unambiguously controllable by computer graphics, with everything on focus at the same apparent distance from the user. Finally, VST systems are much more suited than OST systems, to rendering occlusions between real and virtual elements or to implementing complex visualization processing modalities that are able to perceptually compensate for the loss of the direct real-world view.

Despite all these advantages, the visual perception of the real world with VST HMDs is adversely affected by various geometric aberrations [6–8]. These geometric aberrations are due to the intrinsic features of cameras and displays (e.g., resolutions limitations and optical distortions) and can be boosted by their relative positioning.

One of the major geometric aberrations typical of VST HMDs is related to the misalignment of viewpoints (parallax) between the capturing cameras and the user's perspective through the display (i.e., non-orthoscopic setup). The parallax between capturing camera and user's viewpoint produces distortion into the patterns of horizontal and vertical binocular disparities and this translates into a distorted perception of space.

To recover proper space perception, researchers have put forward various solutions for implementing claimed parallax-free and orthoscopic VST HMDs [9]. In 1998 Fuchs et al. [10] were the first to propose a parallax-free VST HMD. In that system, a pair of mirrors was used to bring the camera centres in the same location of the nodal point of the wearer's eyes.

In a work published in 2000 [11], a systematic analysis of all the possible distortions in depth perception due to non-rigorous orthostereoscopic configurations was presented. Starting from this comprehensive analysis, the authors pursued the same objective of developing a parallax-free VST HMD by means of a set of mirrors and optical prisms whose goal was to align the optical axes of the displays to those of the two cameras. However, also this solution was characterized by a divergence from the conditions of orthostereoscopy in terms of an offset of approximately 30 mm between camera centre and exit pupil of the display, whose effect the authors claimed to be negligible in terms of perceptive distortions of space.

In 2005 State et al. [12], presented an innovative VST HMD specifically designed to generate zero eye-camera offset. Their system, specifically intended for use in medical applications, was designed and optimized through a software simulator the outputs of which then guided the development of a proof-of-concept prototype, built via rapid prototyping and by assembling off-the-shelf components. In their simulated scenario, the authors properly addressed all the aspects for implementing an orthoscopic VST visor; yet their actual embodiment did not satisfy all those requirements due to the constructive complexities (e.g., it did not comprise any eye tracker). Therefore, their system could provide a parallax-free perception of the reality only for user-specific and constant settings in terms of eye position, inter-pupillary distance and eye convergence.

Finally, in 2009 Bottecchia et al. [13], proposed an orthoscopic monocular prototype of VST HMD in which a computer-based correction of the parallax was mentioned. Unfortunately, the authors then did not provide further details on the way the parallax was resolved via software.

Unfortunately, all the presented solutions were bulky and mostly designed for applications in which the pair of stereoscopic cameras is mounted parallel to each other.

By contrast, for those AR applications in which the user is asked to interact with the augmented scene within personal space (i.e., at distances below 2 m), the stereo camera pair ought to be pre-set at a fixed convergence for ensuring sufficient stereo overlaps and granting proper stereo fusion, i.e., toed-in setup [14–17]. This angle of convergence should be established based on assumptions made on the average working distance. In such configuration, for preserving a natural visual perception of the space (i.e., the conditions of orthostereoscopy) and reduce stereoscopic distortions as keystone distortion and depth-plane curvature, theoretically also the two displays should be physically converged of the same angle [11,18]. Yet, this requirement cannot be fulfilled from a practical standpoint and this has implications on the ability of the stereoscopic system to recovering natural depth perception.

When VST HMDs with parallel stereo cameras are intended for use in close-range tasks, a valid alternative is represented by the purely software mechanism proposed in [19]. In their solution, the idea was to maximize via software the stereo overlaps by handling dynamically the convergence or the shearing of the display frustum based on a heuristic estimation of the working distance.

In line with this approach, we here present a method for properly recovering natural stereo fusion of the scene at a predefined distance from the observer in a binocular VST system designed for tasks performed within arm's reach. Our method explicitly takes into consideration the geometry of the setup and the intrinsic parameters of camera and display for computing the appropriate plane-induced homography between the image planes of the stereo cameras and those of their associated displays. On each side, such perspective preserving homography is used for consistently warping the image grabbed by the camera before rendering it onto the corresponding display. This solution, yields a parallax-free perception of the reference plane and, together with the quasi-orthoscopic setup of the VST HMD, manages to recover almost entirely the natural perception of depth in the space around the reference distance. The selection of the reference plane for the homography for a specific use case is based on assumptions made on the average working distance.

For validating our approach, we took advantage of the hybrid nature of a custom-made see-through HMD [20], which supports both video and optical see-through modalities, for drawing an experimental setup whose goal was twofold. First goal was to assess, under OST view, the resulting monoscopic displacement between real features and synthetic ones at various depths around the one taken as reference for the estimation of the homography. The second goal of the tests was to evaluate quantitatively whether and how such displacements affected the perception of the relative depths in the scene under VST view. To this end, we eventually performed preliminary subjective tests aimed at measuring the accuracy in perceiving relative depths through the VST HMD.

2. Materials and Methods

This section is structured as follows. Section 2.1 provides a detailed description of the binocular hybrid video-optical see-through HMD used in this study. Section 2.2 outlines the geometry of the homography induced by a plane that yields a consistent perspective-preserving image warping of the camera frames. Section 2.3 briefly contains a short description of the AR software framework running on the HMD. Finally, Section 2.4 introduces the methodology adopted for validating the method.

2.1. Binocular Hybrid Video/Optical See-Through HMD

In a previous study, we presented a novel approach for the development of stereoscopic AR HMDs able to offer the benefits of both the video and the optical see-through paradigms [20]. The hybrid mechanism was made possible by means of a pair of electrically-driven LC shutters (FOS model by LC-Tec [21]) mounted ahead of the waveguides of a OST HMD, opportunely modified for housing a pair of stereo cameras. The transition between the unaided (OST) and the camera-mediated (VST) view of the real scene is allowed by acting on the transmittance of the electro-optical shutter. As in the first prototype, the hybrid VST/OST HMD is based on a reworked version of a commercial binocular OST HMD (DK-33 by LUMUS [22]). The optical engine of the visor features a 1280 × 720 resolution, a horizontal FoV (hFoV) of 35.2° and a vertical FoV (vFoV) of 20.2°. The stereo camera pair is composed by two Sony FCB-MA13 cameras equipped with a 1/2.45″ CMOS sensor; the cameras are extremely compact in size (16.5 × 10.3 × 18.0 mm) and have the following characteristics: horizontal FoV = 53°, vertical FoV = 29° and frame rate of 30 fps at 1920 × 1080 resolution.

The key differences between the previous prototype and the one that we used in this study are: stereo camera placement and orientation (Figure 1).

As for the stereo camera placement, to pursue a quasi-orthostereoscopic view of the scene under VST modality here we opted for a setup featuring an anthropometric interaxial distance (∼65 mm), hence we mounted the pair of cameras on the top of the two waveguides. To the same end and as previously done in [1,19,23], we opted for a parallel stereo camera setting. Indeed, in AR visors specifically designed for close-up tasks as ours, a toed-in stereo camera setting would undoubtedly widen the area of possible stereo overlaps [17,24]. Yet this configuration, if not coupled with a simultaneous convergence of the optical display axes, would also distort the horizontal and vertical patterns of binocular disparities between the stereo frames. This fact would go against the achievement of a quasi-orthostereoscopic VST HMD and it is deemed to lead to significant distortions in absolute and relative depth perception [25,26].

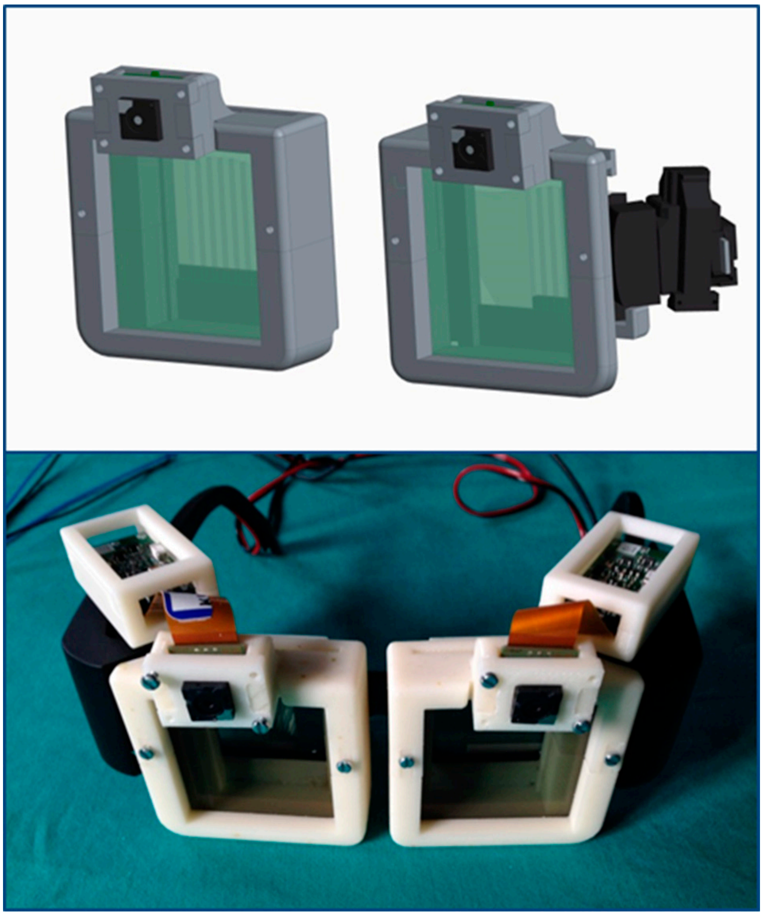

Figure 1. On the top: CAD schematic of the hybrid video/optical see-through head-mounted display comprising the supports for the electro-optical shutters and a pair of stereo cameras mounted on top of the two waveguides of an OST HMD. On the bottom: the HMD.

2.2. *Perspective Preserving Planar Homography*

This section describes the procedure followed for computing the perspective preserving homography. The goal was to find the geometric relation between two perspective visions of a planar scene placed at a pre-defined distance. With reference to Figure 2 and to the equations below, from now on we shall consider the following convention of variables and symbols:

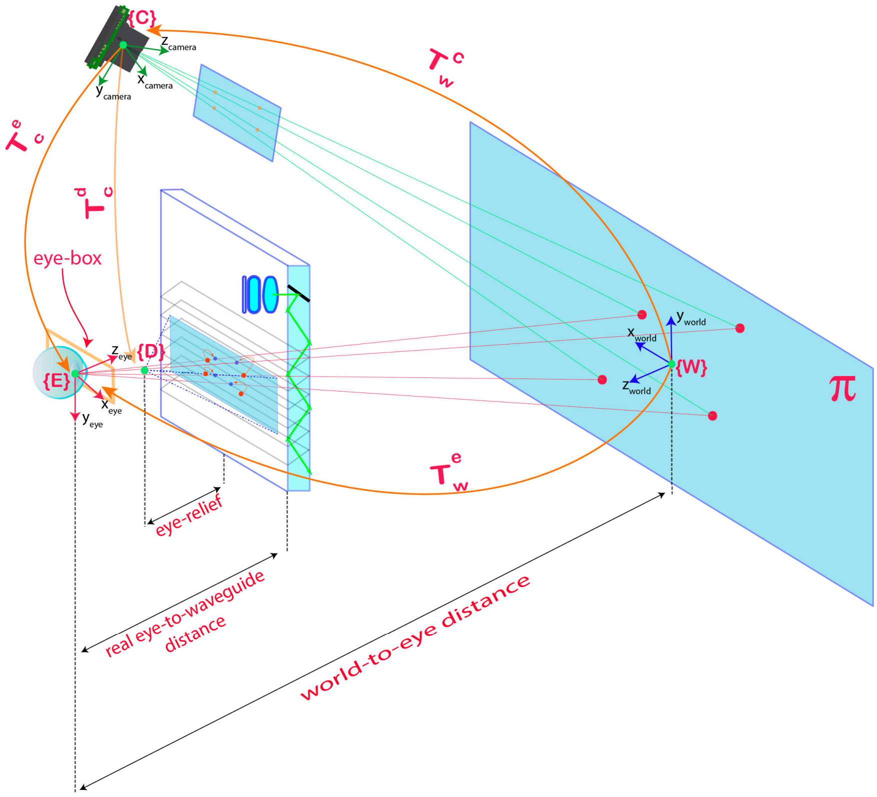

Figure 2. Geometry of the perspective preserving homography induced by a reference plane placed at a pre-defined distance from the eye.

- The homography transformations H_W^D and H_W^C, which relate respectively the points of the reference plane π in the world to their projections onto the image planes of both the display and the camera:

$$\left.\begin{array}{l}\lambda_d x_D = H_W^D\, X_W \\ \lambda_c x_C = H_W^C\, X_W\end{array}\right. \forall\, X_W \in \pi \tag{1}$$

  where the points are expressed in homogeneous coordinates and where λ_c and λ_D are generic scale factors due to the equivalence of homogeneous coordinates rule.
- The distance $d^{D \to \pi}$ between the vertex of the display frustum (D) and the reference plane π.
- The eye relief, which represents the fixed distance between D and the eyepiece lens of the display.
- The eye-box (or eye motion box), which consists of that range of allowed eye's positions, at a pre-established eye-relief distance, from where the full image produced by the eyepiece of the display is visible.
- R_C^D and $\overset{\rightarrow D}{t}_C$, which are respectively the rotation matrix and the translation vector between camera reference system (CRS) and display reference system (DRS).
- K_C and K_D, which are the intrinsic matrixes of camera and display. K_C encapsulates the camera intrinsic parameters and it is computed by using the Zhang's method [27] implemented within the camera calibrator tool of MATLAB. K_D encapsulates the parameters of the near-eye display's frustum and it is approximately derived from the specifics of the HMD as follows [28]. We derived the focal length of the display (f) by using the factory specifics of the horizontal and vertical FoV of the display. In our ideal pinhole camera model of the display, the focal length was set equal on both x-axis and y-axis $(f_x = f_y)$, meaning the pixels of the display were considered as perfectly

square. As coordinates of the principal point (C_u and C_v) we considered exactly half of the display resolution (C_u = Width/2 = 640, C_v = Height/2 = 360). In summary, we assumed:

$$K_D = \begin{bmatrix} f_x & 0 & C_u \\ 0 & f_y & C_v \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} W/\left(2\cdot tan\frac{hFoV}{2}\right) & 0 & W/2 \\ 0 & H/\left(2\cdot tan\frac{vFoV}{2}\right) & H/2 \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

- H_C^D, which is the perspective preserving homography, induced by a fixed plane π placed at distance $d^{D\rightarrow\pi}$ from D.
- $\vec{n}$, which is the normal unit vector to the plane π.

The sought homography transformation H_C^D, describes the point-to-point relation between camera viewpoint and user's viewpoint, such that:

$$\lambda x_D = H_C^D\left(R_C^D, t_C^D,\ K_C, K_D, \pi\right)x_C \tag{3}$$

The parenthesis means that the homography H_C^D is a function of respectively: the relative pose between camera reference system and display reference system (R_C^D, t_C^D), the intrinsic parameters of camera and display (K_C, K_D) and the position and orientation of the reference plane in the scene (π).

For referring everything to the display we can compute H_D^C and inverting the result afterwards (see Equation (4)).

H_C^D is described by a matrix $H_C^D \epsilon\ \Re^{3\times3}$ and it is function of the pose between camera frustum and display frustum and of the two intrinsic matrixes as follows (for referring everything to the display we have computed H_C^D by inverting H_D^C) [29–31]:

$$H_C^D = H_D^{C^{-1}} = \left(K_C\left(R_D^C + \frac{\overset{\rightarrow C}{t}_D\cdot\overset{\rightarrow T}{n}}{d^{D\rightarrow\pi}}\right)K_D{}^{-1}\right)^{-1} \tag{4}$$

The homography transformation (4) is only valid on a fixed plane, perpendicular to the optical axis of the display and placed at a predefined distance ($d^{D\rightarrow\pi}$). If the plane under observation is different or if the observed scene is not planar, its perceived view (through-the-waveguide view) does not match with the rendered image on the display (i.e., direct view and VST view are not orthoscopically registered).

Another important aspect to consider is the actual position of the nodal point of the user's eye (E) with respect to the DRS. This brings about changes in the variables plugged in Equation (3): in the previous equations, we assumed that the nodal point of the eye (eye centre) was located exactly at the vertex of the display frustum (i.e., $E \equiv D$ or $\overset{\rightarrow C}{t}_D \equiv \overset{\rightarrow C}{t}_E$). Unfortunately, this is hardly the case in reality. In addition, we must consider the optical properties of the display eyepiece (i.e., eye relief, eye-box, virtual or focal plane position) (Figure 3). These properties play a role in the way in first approximation the non-ideal eye placement in the display reference system ($E \neq D$) affects the elements of K_D.

In summary, Equation (4) becomes:

$$H_C^E = H_E^{C^{-1}} = \left(K_C\left(R_D^C + \frac{\overset{\rightarrow C}{t}_E\cdot\overset{\rightarrow T}{n}}{d^{E\rightarrow\pi}}\right)\widetilde{K_D}^{-1}\right)^{-1} \tag{5}$$

where $\overset{\rightarrow C}{t}_E$ ought to be known and where the intrinsic matrix of the display $\widetilde{K_D}$ is different from the original K_D. In view of these considerations, an orthoscopic alignment is attained in theory only if we could determine with absolute accuracy the user's eyes position in the HMD's eyepiece reference system (i.e., DRS). Indeed, in Equation (3) the pose between eye and camera assumes a key role.

Unfortunately, the eye position in the DRS and consequently in the CRS, varies according to how the HMD is worn and is dependent on the user's facial shape (e.g., inter-pupillary distance).

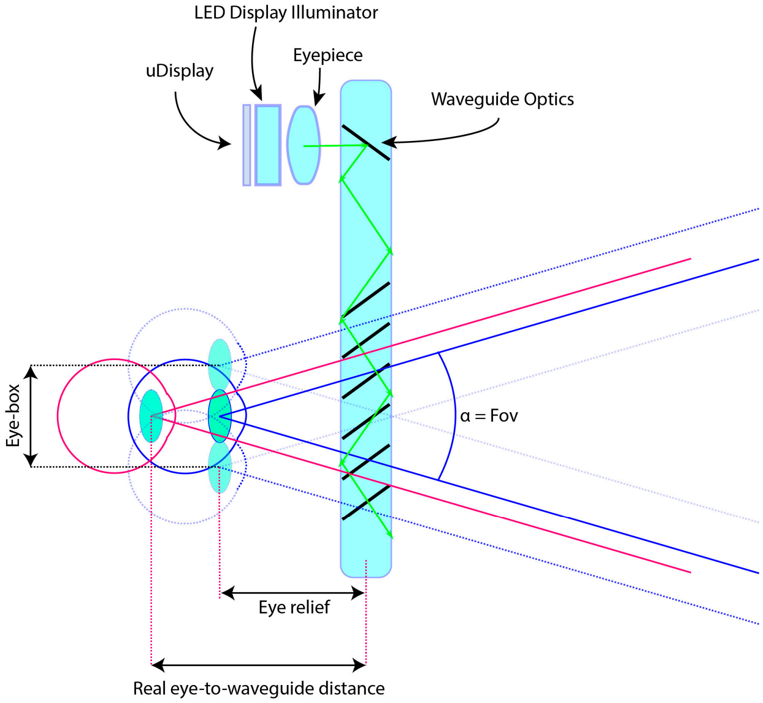

Figure 3. Optical properties of the near-eye display.

Since our HMD did not comprise any eye-tracker to calibrate the stereo cameras to the user's inter-pupillary distance and since we did not perform any specific display calibration (for precisely determining $\widetilde{K_D}$), in our tests we determined an approximation of the homography $\widetilde{H_C^E}$ as follows. We asked the user to wear the HMD and observe under OST modality a target checkerboard placed orthogonally to his own viewpoint at approximately the distance of the homography plane π (Figure 4). We then performed, in real time, multiple refinements of the initial homography H_C^D by means of an additional translational homography $(\widetilde{h})$ whose role was to align the user's views of the checkerboard (real and synthetic). With this additional homography we intended to compensate for the uncertainties in defining the actual position of the eye E with respect to D. In our method, we only considered the effect of translational movements along the x and y-axis (parallel to the image plane). Thus, to a first approximation, we excluded the effect of a non-perfect placement of the user's eye centres at the eye-relief (i.e., we assumed: eye relief = real eye-to-waveguide distance). The relation between the approximated homography $\widetilde{H_C^E}$ and the ideal H_C^D then becomes:

$$\widetilde{H_C^E} = \widetilde{h}\, H_C^D = \begin{bmatrix} 1 & 0 & x_p \\ 0 & 1 & y_p \\ 0 & 0 & 1 \end{bmatrix} H_C^D \qquad (6)$$

In conclusion, unlike the method proposed by Tomioka et al. [30], we estimate the user-specific homography uniquely by means of design and calibration data. The homography is then refined to embody the effects of the intrinsic parameters of the displays and of the non-ideal eyes placement in the HMD so as to be adapted to the subject's interpupillary distance (IPD).

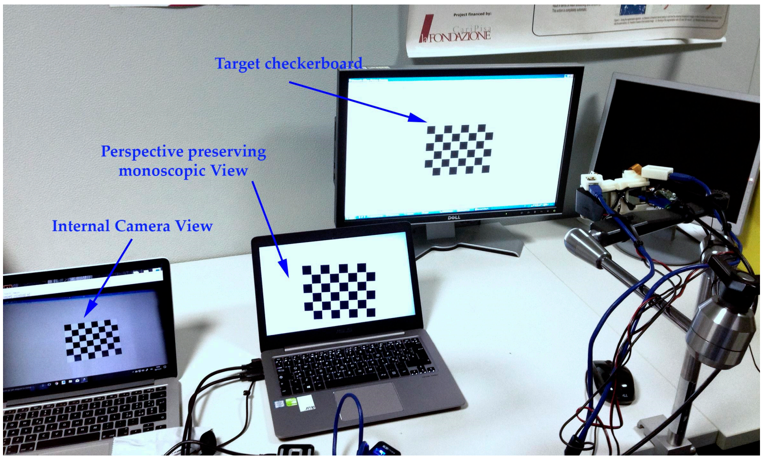

Figure 4. Experimental setup for measuring the monoscopic disparities. The same target checkerboard was used also for the user-specific homography refinement.

2.3. Software Application

For validating our method, we developed a dedicated software application whose main goal was to manage the camera frames as follows. Camera frames are first grabbed and opportunely undistorted for eliminating the non-linearities due to radial distortions. The undistorted frames are warped according to the perspective preserving homography. The warped frames are rendered onto the background of a stereoscopic scenegraph that is finally screened onto the binocular HMD. For this application, we did not add any properly registered virtual content to the scenegraph, as our objective was uniquely to perform perceptual studies on how depth perception was retrieved under VST view.

The application was created in the form of a single executable file with shared libraries all built in C++, following the same logic of the AR software framework previously developed in [32]. As for the library managing the rendering of the scenegraph, we used the open-source library for 3D computer graphics and visualization Visualization Toolkit (VTK), version 7.0.0 [33]. As regards the machine vision routines, needed for processing the camera frames before rendering them onto the background of the scene-graph, we adopted the open-source software library OpenCV 3.1 [34].

2.4. Tests

The proposed solution combines a perspective preserving warping mechanism with a quasi-orthoscopic setting of the VST HMD. The goal of our tests was to assess the efficacy of such solution in recovering the natural perception of depth in the space around a pre-established distance from the observer. We grouped the tests into two basic categories: tests for measuring the patterns of on-image disparities, under OST modality, between real features and HMD-mediated ones and tests for assessing objectively and subjectively the depth estimation accuracy under VST modality.

2.4.1. Test 1: Measure of on-Image Displacements between Direct View and VST View

For measuring the patterns of monoscopic disparities between direct view and VST view, we used the experimental setup showed on Figure 4. The on-image displacements between real features and HMD-mediated ones were measured by means of an additional video camera (Sony FCB-MA13) placed approximately at the ideal eye's position (internal camera). As target scene, we used a standard checkerboard of size 160 × 120 mm (with square size 20 mm) that was displayed on an external monitor. The internal camera was able to capture two views of the target scene: a direct view and a VST view (Figure 5). The corners of the checkerboard could be robustly detected through a Matlab's function for corners detection. The on-image displacements or monoscopic disparities between the image coordinates of the real and VST views of the corner points were so easily determined. The real poses of the target planes with respect to the internal camera were estimated by solving a standard perspective-n-point problem and knowing the 3D-2D point correspondences.

We repeated such measurements at various depths around the one taken as reference for the estimation of the homography (plane π). The range of depths for which the disparities were measured was: (250–650) mm.

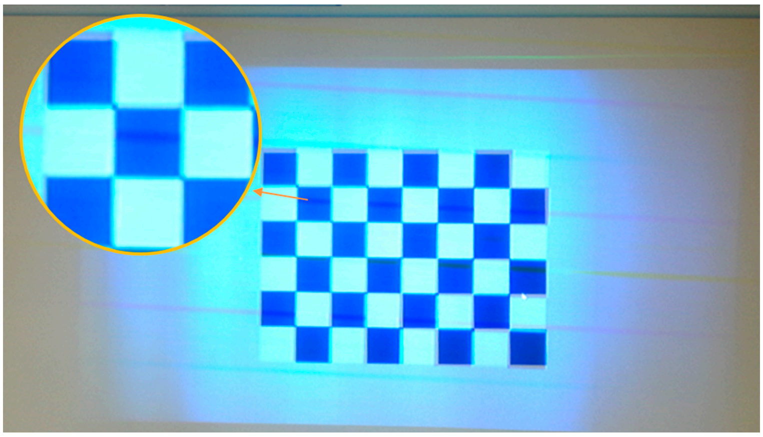

Figure 5. On-image displacements between direct view and VST view of a target scene. The test images were grabbed by an additional video camera placed approximately at the eye relief point of the eyepiece of the HMD.

2.4.2. Test 2: Assessment of Depth Perception through Objective and Subjective Measures

For assessing the degree of accuracy in depth estimation under VST view, we conducted 2 different sets of measurements. At first, we computed the resulting angles of retinal disparities yielded by the monoscopic displacements on both views; these binocular disparities can provide a quantitative estimation of the uncertainty in detecting the relative depths between objects due to the non-ideally orthoscopic setting of the VST HMD. Secondly, we performed a preliminary user study aimed at assessing the accuracy in perceiving relative depths at different distances within personal space (within 1.2 m). Before the session of tests, the homography was refined under OST modality to be adapted to the subject's IPD.

In the tests, one participant wearing the HMD under VST modality was asked to estimate the relative depth relations between three objects of same size and colour (three yellow Lego® bricks of size 9.6 × 32 × 16 mm). We engaged only one participant so as to be consistent in terms of user's stereoscopic acuity. The bricks were laid on five different A3 paper sheets (size 297 × 420 mm), each of which provided with demarcation lines indicating different relative depths. In each triplet of demarcation lines, the relative distances between the bricks were decided randomly, with a defined relative distance between two adjacent bricks of 2 mm.

The paper sheets were placed at five different distances from the observer (Figure 6), covering a range of depths of about 900 mm (i.e., from a minimum absolute depth of 300 mm to a maximum depth of 1200 mm). For each position of the paper sheet, the test was repeated four times.

The perceptual tests were all performed keeping the same homography transformation. For all the target planes, the computed homography was referred to a reference plane perpendicular to the optical axis of the display and placed at a distance of 500 mm. The final goal of the tests was indeed to assess on how this aspect would have had a detrimental effect on perceiving relative depths for all the tested distances of the triplets of bricks.

Each paper included ten possible configurations of relative positions between the three bricks, so we tested a total of 4 (n° of sessions) × 5 (n° of paper sheets at different depths) × 10 (n° of configurations of triplets per paper) = 200 configurations of triplets of bricks.

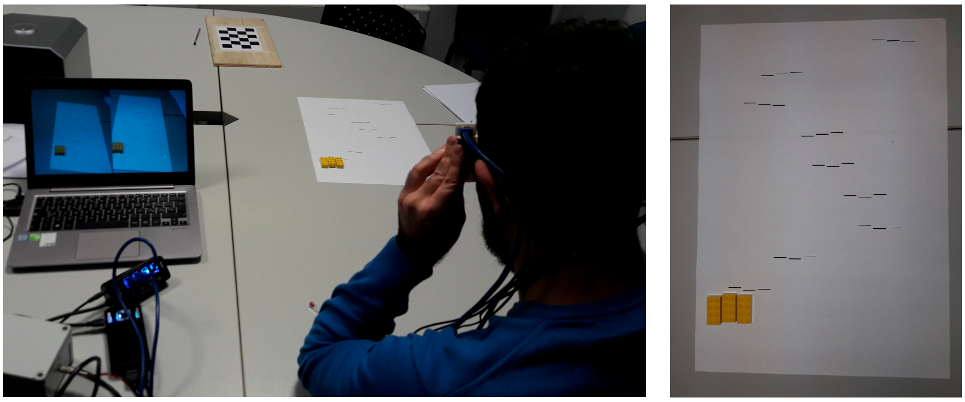

Figure 6. On the left: experimental setup for depth estimation tests. On the right: paper sheet with marked segments. The segments were used for placing the triplets of bricks at different absolute depths from the observer and with random patterns of relative depths among them.

3. Results

This section reports on the results of the two sessions of tests.

3.1. Results of Test 1

We measured the patterns of monoscopic disparities by moving the target checkerboard at ten different positions with respect to the HMD. Each checkerboard contains a set of 35 corner points which results in a total of 350 feature points to be considered in our evaluation. In Figures 7 and 8, the resulting horizontal (h^d) and vertical (v^d) disparities for all the ten positions of the target plane are shown in function of the z-coordinate of the point. The z coordinate of the point is retrieved knowing the pose of the target plane to which they belong. The maps of disparities for each target position are reported in Appendix A.

In relation to the distance from the reference plane π, the vertical disparities follow a steeper increase with respect to the horizontal disparities. This fact directly results from the vertical parallax between CRS and DRS, while the horizontal disparities are only functions of the distance from the reference plane. In Figure 9, we show the horizontal disparities for the points belonging to the six target planes closer to the reference plane. Here the range of depths is: (479–555) mm. In Table 1 the values of the mean and standard deviation of the horizontal disparities are reported for all the target positions.

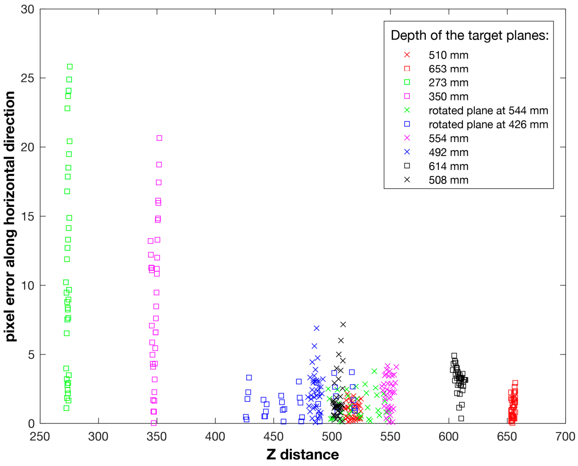

Figure 7. Horizontal monoscopic disparities.

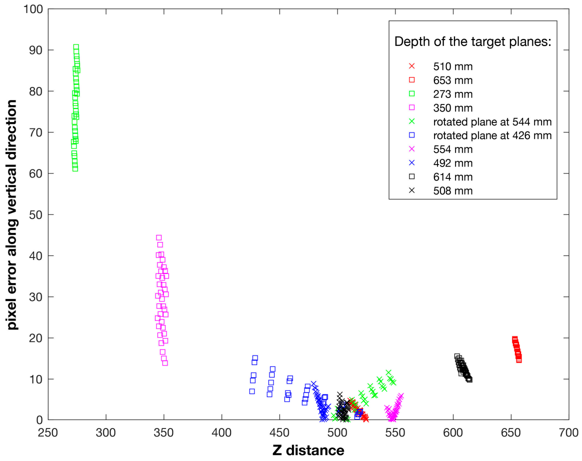

Figure 8. Vertical monoscopic disparities.

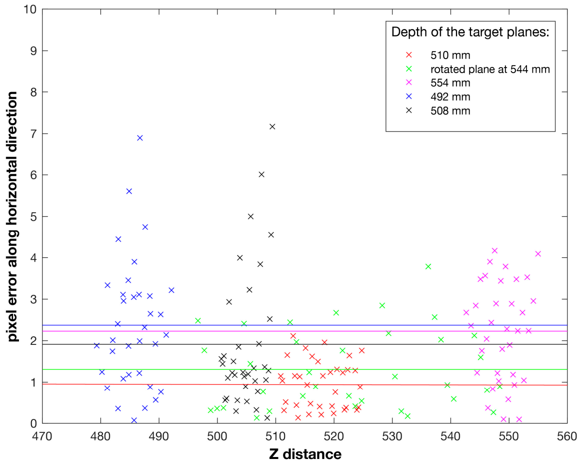

Figure 9. Horizontal monoscopic disparities for the target planes around z = 500 mm. The coloured lines are the mean pixel errors on the u coordinate for each target plane.

Table 1. Mean and standard deviation of horizontal monoscopic disparities between direct-view and VST views of the target planes.

Target Plane Position	Mean Error on u Coordinate (Pixel)	Error’s Standard Deviation
273 mm	11.2072	7.6836
350 mm	8.8623	5.6106
426 mm (rotated)	1.5173	1.0251
492 mm	2.3730	1.5282
508 mm	1.9128	1.6844
510 mm	0.9542	0.5848
544 mm (rotated)	1.3085	0.9572
554 mm	2.2297	1.1904
614 mm	3.1658	0.9445
653 mm	1.1834	0.7813

3.2. *Results of Test 2*

3.2.1. Estimation of Depth Perception under VST View

We here provide a quantitative estimation of the misperception of depth due to the unwanted disparities on both the sides of the HMD. In stereoscopic displays as binocular HMDs, human stereopsis is usually simulated by generating pairs of stereo views with a perceptually consistent amount of horizontal disparities. The relation between the binocular horizontal disparity observed on the 3D display (d_d) and the retinal disparity (d_r) is the following [35]:

$$d_r = 2 \cdot atan\left(\frac{d_d \cdot tan(\alpha/2)}{W}\right) \approx \frac{2 \cdot d_d \cdot tan(\alpha/2)}{W} \tag{7}$$

By plugging $W = 1280$ and $\alpha = 35.2°$ into the equation, we can calculate the minimum angular disparity (or minimum angular resolution) that our HMD is able to provide (for a display disparity $d_d = 1$ pixel): $d_{rmin} \cong 1.7$ arcmin. This value leads to a visual acuity of about half of the average visual acuity in human vision (visual acuity $= {d_r}^{-1}$). The approximated formula of the depth resolution dZ at a distance Z from the observer can be retrieved as follows [18,36,37]:

$$dZ = \frac{Z^2 d_r}{I \cdot K_r} \tag{8}$$

where d_r is expressed in arcmins, I is the observer's IPD and $K_r = 3437.75$ is a constant that relates radians to arcmins. By considering a standard value of $I = 65$ mm, the stereoacuity or depth resolution offered by our HMD at 500 mm is of about 2 mm.

By plugging Equation (8) in Equation (7), we obtain the relation between depth resolution and binocular horizontal disparity:

$$dZ = \frac{2 \cdot Z^2 \cdot d_d \cdot tan(\alpha/2)}{W \cdot I \cdot K_r} \tag{9}$$

where the binocular horizontal disparity can be expressed in terms of image coordinates as follows: $d_d = u_r - u_l$.

If we consider the values of the stereoacuity offered by our HMD at different depths, we are able to compute the ideal density of the homographies from Z_{min} to Z_{max}:

$$Z_i = Z_{i-1} + dZ_i = Z_{i-1} + \frac{{Z_i}^2 d_r}{I \cdot K_r} \tag{10}$$

For instance, in the range of depths between 250 and 650 mm ($Z_0 = 250$ and $Z_{max} \geq 650$), in theory we would need as much as 350 different homographies in order to stay within the resolution constraints of the HMD. In spite of this and as we explained in Section 2.4.2, the perceptual tests were all performed keeping the same homography transformation for all the target planes, since our goal was to assess on how this aspect would have affected depth perception.

In the first session of tests, we observed how the non-ideally orthoscopic setting of the HMD causes unwanted monoscopic disparities h^d on both the sides of the HMD the further we go from the reference distance. Thus, the horizontal disparity can be written as follows:

$$\widetilde{d_d} = u_r \pm \left(h_r^d\right) - \left(u_l \pm h_l^d\right) = d_d \pm 2 \cdot h^d \tag{11}$$

In the equation, we assumed the worst-case scenario, where monocular disparities on both sides add together and they have the same value. In this way, we can estimate the contribution of such disparities to the depth resolution:

$$\widetilde{dZ} = \frac{2 \cdot Z^2 \cdot \left(d_d + 2 \cdot h^d\right) \cdot tan(\alpha/2)}{W \cdot I \cdot K_r} \tag{12}$$

So, the overall depth resolution is affected by the additional disparity contribution brought by the non-ideally orthoscopic setting of the visor. In the range of depths around the plane π (z = 500 mm) used for computing the homography, the mean of $\widetilde{dZ}$ was of about 8.8 mm. The value of $\widetilde{dZ}$ is lower if we consider a smaller area at the center of the stereo images, that is where the monoscopic horizontal disparities are not as high.

In Figure 10, the profile of $\widetilde{dZ}$ at various distances from the observer is shown for different values of horizontal disparities. In the figure, we also report the values of $\widetilde{dZ}$ associated to the measured values of h^d for $Z \in [490 - 510]$ mm.

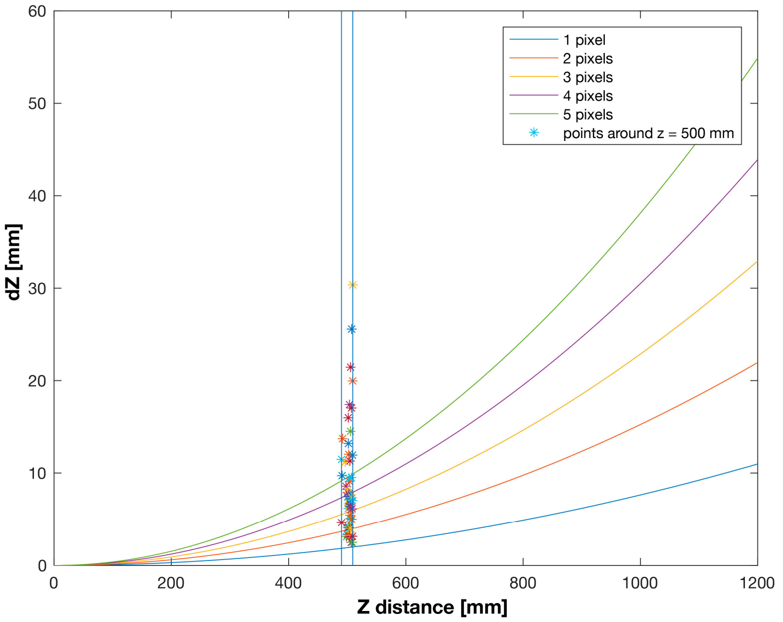

Figure 10. Depth resolution vs distance from the observer with different values of horizontal disparities. The asterisks are the values of $\widetilde{dZ}$ associated to the measured disparities around plane π.

3.2.2. Perceptual Tests

In Table 2, the results of the perceptual tests are reported. We measured the rate of success in terms of proper relative depths estimation among the bricks in all the tested configurations.

The success rate was surprisingly higher than expected (98.5% of total success rate), taking into consideration the fact that the relative depths between the three bricks was between 2 and 4 mm at any distance. In the next session, we shall motivate for the apparent inconsistency between the misperception of relative depths predicted by the mathematical models and the perceived distortion of space experienced by the user during the real use cases.

Table 2. Results of the depth estimation tests.

Distance Range (mm)	**First Test Success Rate**	**Second Test Success Rate**	**Third Test Success Rate**	**Fourth Test Success Rate**	
~300–~750	100%	100%	100%	100%	
~400–~850	100%	100%	100%	100%	
~500–~950	100%	90%	100%	100%	
~600–~1050	100%	100%	90%	100%	
~700–~1150	100%	100%	90%	100%	
Total Success Rate for each test	100%	98%	96%	100%	**Total Success Rate** 98.5%

4. Discussion & Conclusions

In this work, we have presented a VST HMD whose geometry violates the rigorous conditions of orthostereoscopy. For properly recovering natural stereo fusion of the scene in a region around a predefined distance from the user, we partially resolve the eye-camera parallax by warping the camera images through a perspective preserving homography.

The appropriate plane-induced homography between the image planes of the pair of stereo cameras and those of their associated displays, is computed by explicitly taking into consideration the geometry of the VST HMD and the intrinsic parameters of camera and display. The homography is therefore estimated uniquely by means of design and calibration data.

For validating our solution, we conducted objective and subjective tests. The goal of the tests was to assess the efficacy of such solution in recovering the natural perception of depth in the space around a pre-established distance from the observer.

Thanks to the hybrid nature of the HMD, which can work also under OST modality, in the first session of tests we measured the patterns of on-image disparities between a direct view of the world and a VST view. An internal camera, placed at the ideal eye's position, captured both the views of a target plane at different distances and orientations relative to the HMD.

These monoscopic disparities provided an initial measure of the amount of perceptual distortions brought by the non-orthoscopic setting of the HMD. The same disparities were then used to quantitatively estimate the resulting degree of uncertainty in perceiving relative depths under VST view.

Finally, we performed subjective tests aimed at assessing under real-use conditions, the actual depth estimation accuracy under VST view.

From a human-factor standpoint, VST HMDs raise issues related to the user's interaction with the augmented content and to some perceptual conflicts. With stereoscopic VST HMDs, the user can perceive relative depths between real and/or virtual objects by providing consistent depth cues in the recorded images delivered to the left and right eyes by the two displays of the visor. In our tests, we focused on relative depth measurements since relative depths information are much more important than absolute depths for aiding manual tasks in the personal (and intimate) space [7,8].

However, depth perception in binocular VST HMDs has not been fully investigated in literature. In their study, Kyto et al. [7] performed perceptual tests with a stereoscopic VST HMD aimed at measuring the effect of binocular disparities, relative size and height in the visual field on depth judgments in the action space (distances from 2 to 20 m). Their main finding was that depth perception through VST view in the action space is highly improved by a proper combination of a virtual content (i.e., auxiliary augmentations) providing binocular disparity and relative size cues. In our study, we did not use any sort of auxiliary augmentation since the goal of our perceptual studies was to assess how depth perception is recovered when using non-orthoscopic VST HMDs. Further, our depth judgment tests were performed within the personal and intimate space where the visual interaction with the augmented scene is likely to hide the ground plane and for which other depth cues other than binocular disparities and occlusions are not as relevant.

Overall, the obtained results were surprisingly positive in terms of depth judgment tasks. This is in line with what experienced by State et al. [19] and suggested by Milgram et al. [38], who both asserted that the distortion of the visual space derived from the mathematical models underpinning stereo vision is significantly higher than what the user perceives in reality. In our opinion, this fact is mostly motivated by the presence of other binocular depth cues as eyes convergence or monocular ones as linear perspective, texture gradient, shades and shadows [39]. All these cues contribute to provide a finer perception of depth in the personal space and partially compensate for the distortions brought by the non-orthoscopic setting of the VST HMD. The results of our preliminary perceptual tests were even more positive than the ones presented in [19], as in our tests the user could not use the hand as a "visual aid" for relative and absolute distance estimations.

Another aspect to consider is that the distortion of the patterns of binocular horizontal disparities at distances different from the homography plane, is not as severe at the centre of the stereo images, which is where the user normally directs his own view. Further, even if the vertical monoscopic disparities follow a steeper trend as the distance from the homography plane increases, we believe that their effect onto the perception of relative depths is not as evident, also considering that their combinatory effect is likely to be null.

In conclusion, the quasi-orthoscopic setting of the HMD and the user-specific homography, refined to embody the effects of the intrinsic parameters of the displays and of the non-ideal eyes placement in the HMD, are sufficient to recover a proper perception of relative depths in the personal space. Further, we can assert that the actual density of homographies that ensures a non-perceptible distortion of the visual space in the personal space can be sparser than the ideal pattern retrieved by estimating the trend of the stereoacuities of the HMD.

All of this suggests that we should investigate whether display calibration and eye-tracking can allow the achievement of similar results without the need for a user-specific homography refinement. Display calibration would in fact provide a proper estimation of the linear and non-linear projective parameters of the display, while eye-tracking would yield a robust and reliable estimation of the eyes position.

It is important to outline that the results of our perceptual tests can be considered as a preliminary proof of effectiveness of the proposed solution in recovering natural depth perception in a quasi-orthoscopic VST HMD. In addition, the testing platform herein used strongly encourages us to conducting structured user-studies involving more subjects and aimed at investigating further on how our solution can be of help to the VR and AR communities for investigations relative to user's perception and task achievement efficiency, hence in fields as human-computer interaction, neuroscience and human factor in computing systems.

Acknowledgments: This work was funded by the HORIZON2020 Project VOSTARS, Project ID: 731974. Call: ICT-29-2016—Photonics KET 2016.

Author Contributions: Fabrizio Cutolo, Umberto Fontana and Vincenzo Ferrari designed and developed the head-mounted display. Fabrizio Cutolo elaborated the proper scientific framework of the proposed solution and explained the differences between this paper and previous solutions in the field. Fabrizio Cutolo, Umberto Fontana and Vincenzo Ferrari conducted the tests and Fabrizio Cutolo and Umberto Fontana analyzed and discussed the results.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Hereafter the disparity maps for all the target planes considered are shown.

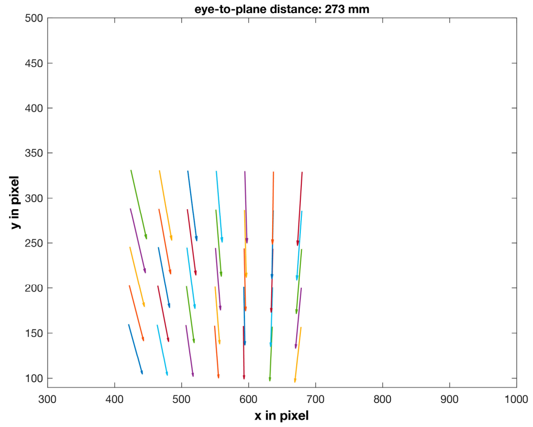

Figure A1. Map of disparities for a target plane placed at 273 mm from the observer.

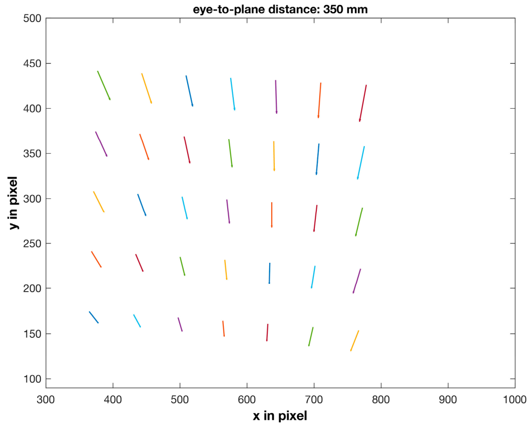

Figure A2. Map of disparities for a target plane placed at 350 mm from the observer.

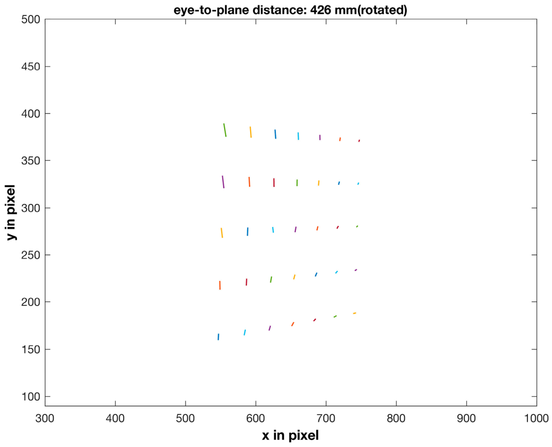

Figure A3. Map of disparities for a target plane rotated and placed at 426 mm from the observer.

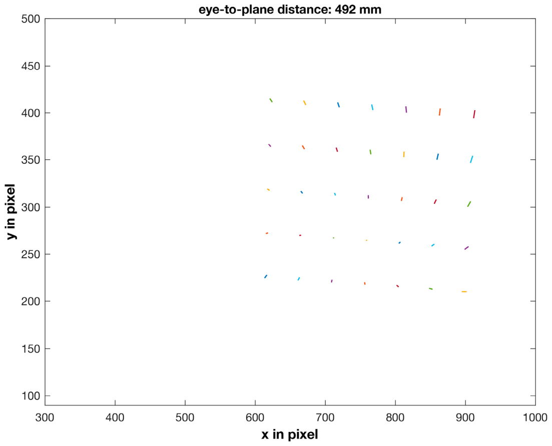

Figure A4. Map of disparities for a target plane placed at 492 mm from the observer.

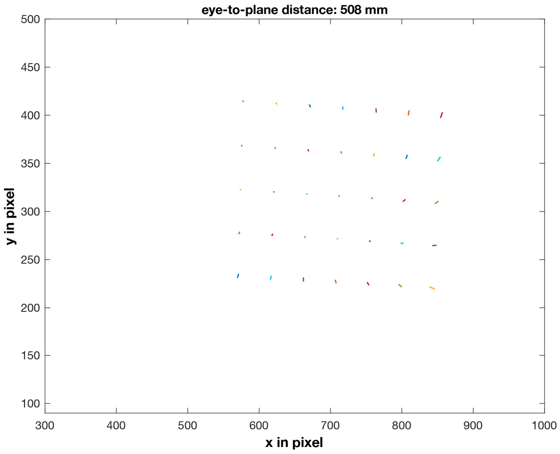

Figure A5. Map of disparities for a target plane placed at 508 mm from the observer.

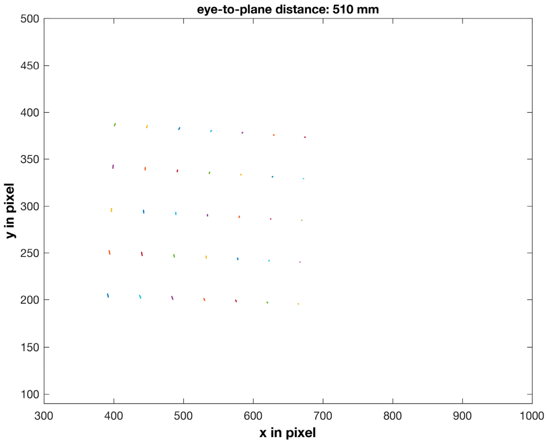

Figure A6. Map of disparities for a target plane placed at 510 mm from the observer.

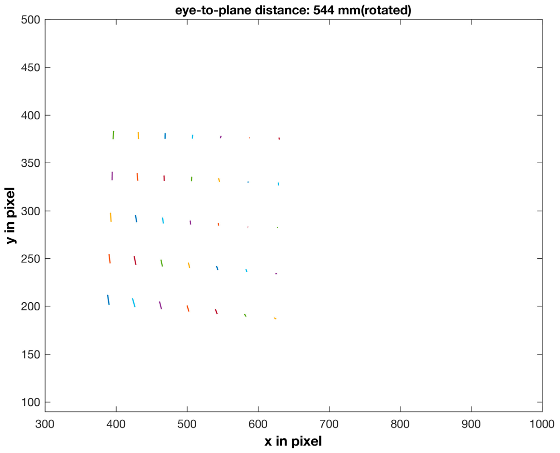

Figure A7. Map of disparities for a target plane rotated and placed at 544 mm from the observer.

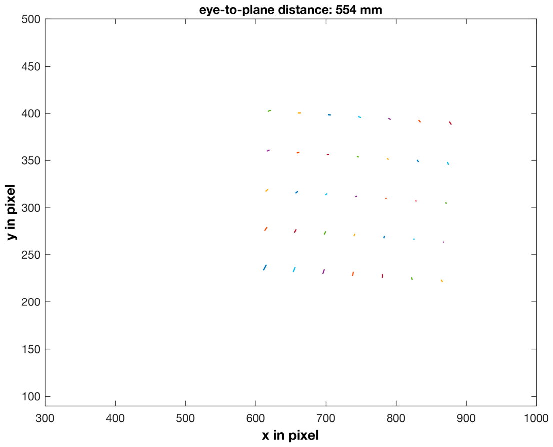

Figure A8. Map of disparities for a target plane placed at 554 mm from the observer.

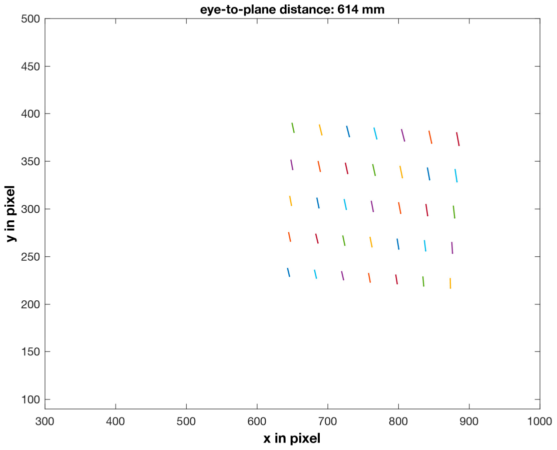

Figure A9. Map of disparities for a target plane placed at 614 mm from the observer.

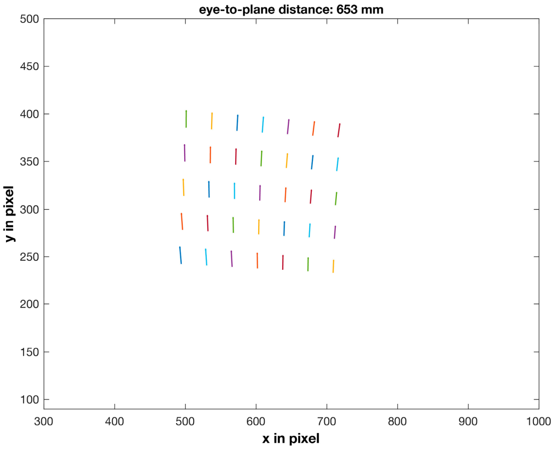

Figure A10. Map of disparities for a target plane placed at 653 mm from the observer.

References

1. Cutolo, F.; Freschi, C.; Mascioli, S.; Parchi, P.; Ferrari, M.; Ferrari, V. Robust and accurate algorithm for wearable stereoscopic augmented reality with three indistinguishable markers. *Electronics* **2016**, *5*, 59. [CrossRef]

2. Rolland, J.P.; Fuchs, H. Optical versus video see-through mead-mounted displays in medical visualization. *Presence Teleoper. Virtual Environ.* **2000**, *9*, 287–309. [CrossRef]
3. Benton, S.A. *Selected Papers on Three-Dimensional Displays*; SPIE Optical Engineering Press: Bellingham, WA, USA, 2001.
4. Rolland, J.P.; Holloway, R.L.; Fuchs, H. A comparison of optical and video see-through head-mounted displays. *Telemanip. Telepresence Technol.* **1994**, *2351*, 293–307.
5. Cutolo, F. Augmented reality in image-guided surgery. In *Encyclopedia of Computer Graphics and Games*; Lee, N., Ed.; Springer International Publishing: Cham, Switzerland, 2017; pp. 1–11.
6. Kytö, M. *Depth Perception of Augmented and Natural Scenes through Stereoscopic Systems*; Aalto University: Espoo, Finland, 2014.
7. Kyto, M.; Makinen, A.; Tossavainen, T.; Oittinen, P. Stereoscopic depth perception in video see-through augmented reality within action space. *J. Electron. Imaging* **2014**, *23*, 011006. [CrossRef]
8. Cutting, J.E.; Vishton, P.M. Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In *Perception of Space and Motion*; Academic Press: San Diego, CA, USA, 1995; pp. 69–117.
9. Drascic, D.; Milgram, P. Perceptual issues in augmented reality. In Proceedings of the SPIE International Society for Optical Engineering, San Jose, CA, USA, 31 January–1 February 1996; pp. 123–134.
10. Fuchs, H.; Livingston, M.A.; Raskar, R.; Colucci, D.; Keller, K.; State, A.; Crawford, J.R.; Rademacher, P.; Drake, S.H.; Meyer, A.A. Augmented reality visualization for laparoscopic surgery. In Proceedings of the First International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Cambridge, MA, USA, 11–13 October 1998; pp. 934–943.
11. Takagi, A.; Yamazaki, S.; Saito, Y.; Taniguchi, N. Development of a stereo video see-through hmd for ar systems. In Proceedings of the IEEE and ACM International Symposium on Augmented Reality, Munich, Germany, 5–6 October 2000; pp. 68–77.
12. State, A.; Keller, K.P.; Fuchs, H. Simulation-based design and rapid prototyping of a parallax-free, orthoscopic video see-through head-mounted display. In Proceedings of the International Symposium on Mixed and Augmented Reality, Vienna, Austria, 5–8 October 2005; pp. 28–31.
13. Bottecchia, S.; Cieutat, J.-M.; Merlo, C.; Jessel, J.-P. A new ar interaction paradigm for collaborative teleassistance system: The POA. *Int. J. Interact. Des. Manuf.* **2009**, *3*, 35–40. [CrossRef]
14. Livingston, M.A.; Ai, Z.M.; Decker, J.W. A user study towards understanding stereo perception in head-worn augmented reality displays. In Proceedings of the 8th IEEE International Symposium on Mixed and Augmented Reality—Science and Technology, Orlando, FL, USA, 19–22 October 2009; pp. 53–56.
15. Matsunaga, K.; Yamamoto, T.; Shidoji, K.; Matsuki, Y. The effect of the ratio difference of overlapped areas of stereoscopic images on each eye in a teleoperation. *Proc. Spiel. Int. Soc. Opt. Eng.* **2000**, *3957*, 236–243.
16. Woods, A.; Docherty, T.; Koch, R. Image distortions in stereoscopic video systems. *Stereosc. Disp. Appl. IV* **1993**, *1915*, 36–48.
17. Ferrari, V.; Cutolo, F.; Calabrò, E.M.; Ferrari, M. [Poster] HMD video see though AR with unfixed cameras vergence. In Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Munich, Germany, 10–12 September 2014; pp. 265–266.
18. Cutolo, F.; Ferrari, V. The role of camera convergence in stereoscopic video see-through augmented reality displays. In Proceedings of the Future Technologies Conference (FTC), Vancouver, BC, Canada, 29–30 November 2017; pp. 295–300.
19. State, A.; Ackerman, J.; Hirota, G.; Lee, J.; Fuchs, H. Dynamic virtual convergence for video see-through head-mounted displays: Maintaining maximum stereo overlap throughout a close-range work space. In Proceedings of the IEEE and ACM International Symposium on Augmented Reality, New York, NY, USA, 29–30 October 2001; pp. 137–146.
20. Cutolo, F.; Fontana, U.; Carbone, M.; Amato, R.D.; Ferrari, V. [Poster] hybrid video/optical see-through HMD. In Proceedings of the 2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct), Nantes, France, 9–13 October 2017; pp. 52–57.
21. LC-TEC Advanced Liquid Crystal Optics. Available online: http://www.lc-tec.se/ (accessed on 30 October 2017).
22. Lumus. Available online: http://lumusvision.com (accessed on 30 October 2017).

23. Kanbara, M.; Okuma, T.; Takemura, H.; Yokoya, N. A stereoscopic video see-through augmented reality system based on real-time vision-based registration. In Proceedings of the IEEE Virtual Reality 2000 (Cat. No. 00CB37048), New Brunswick, NJ, USA, 18–22 March 2000; pp. 255–262.
24. Allison, R.S. Analysis of the influence of vertical disparities arising in toed-in stereoscopic cameras. *J. Imaging Sci. Technol.* **2007**, *51*, 317–327. [CrossRef]
25. Banks, M.S.; Read, J.C.A.; Allison, R.S.; Watt, S.J. Stereoscopy and the human visual system. *SMPTE Motion Imaging J.* **2012**, *121*, 24–43. [CrossRef] [PubMed]
26. Vienne, C.; Plantier, J.; Neveu, P.; Priot, A.E. The role of vertical disparity in distance and depth perception as revealed by different stereo-camera configurations. *I-Perception* **2016**, *7*. [CrossRef] [PubMed]
27. Zhang, Z.Y. A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal.* **2000**, *22*, 1330–1334. [CrossRef]
28. Grubert, J.; Itoh, Y.; Moser, K.R.; Swan, J.E., II. A survey of calibration methods for optical see-through head-mounted displays. *IEEE Trans. Vis. Comput. Graph.* **2017**. [CrossRef] [PubMed]
29. Hartley, R.; Zisserman, A. *Multiple View Geometry in Computer Vision*; Cambridge University Press: Cambridge, UK, 2003.
30. Tomioka, M.; Ikeda, S.; Sato, K. Approximated user-perspective rendering in tablet-based augmented reality. In Proceedings of the 2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)—Science and Technology, Adelaide, Australia, 1–4 October 2013; pp. 21–28.
31. Lothe, P.; Bourgeois, S.; Royer, E.; Dhome, M.; Naudet-Collette, S. Real-time vehicle global localisation with a single camera in dense urban areas: Exploitation of coarse 3D city models. In Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 863–870.
32. Cutolo, F.; Siesto, M.; Mascioli, S.; Freschi, C.; Ferrari, M.; Ferrari, V. Configurable software framework for 2D/3D video see-through displays in medical applications. In *Augmented Reality, Virtual Reality, and Computer Graphics, Part II, Proceeding of the Third International Conference, AVR 2016, Lecce, Italy, 15–18 June 2016*; De Paolis, L.T., Mongelli, A., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 30–42.
33. Vtk Visualization Toolkit. Available online: https://www.vtk.org/ (accessed on 30 October 2017).
34. Opencv. Open Source Computer Vision Library. Available online: https://opencv.org/ (accessed on 30 October 2017).
35. Gadia, D.; Garipoli, G.; Bonanomi, C.; Albani, L.; Rizzi, A. Assessing stereo blindness and stereo acuity on digital displays. *Displays* **2014**, *35*, 206–212. [CrossRef]
36. Harris, J.M. Monocular zones in stereoscopic scenes: A useful source of information for human binocular vision? In *Stereoscopic Displays and Applications XXI*; SPIE: Bellingham, WA, USA, 2010; pp. 151–162.
37. Kyto, M.; Nuutinen, M.; Oittinen, P. Method for measuring stereo camera depth accuracy based on stereoscopic vision. In *Three-Dimensional Imaging, Interaction, and Measurement*; SPIE: Bellingham, WA, USA, 2011.
38. Milgram, P.; Kruger, M. Adaptation effects in stereo due to online changes in camera configuration. In *Stereoscopic Displays and Applications III*; SPIE: Bellingham, WA, USA, 1992; pp. 122–134.
39. Tovée, M.J. *An Introduction to the Visual System*; Cambridge University Press: Cambridge, UK, 1996.