3dvisionlabs

WHITE PAPER

Beginners Guide to Omnidirectional Stereo Vision

The advantages of ultra-wide (omnidirectional) versus rectilinear stereo vision and the science behind the HemiStereo® technology.

AUTHORS

Michel Findeisen
Co-Founder & CEO

Lars Meinel
Co-Founder & CEO

SUMMARY

This white paper briefly reviews the major concepts and characteristics of conventional rectilinear stereo vision in the context of Intelligent Video Analytics (IVA). The text further conveys the advantages of ultra-wide 3D vision using fish-eye lenses. The presented scientific concepts form the foundation of the HemiStereo® depth sensing technology.

INTRODUCTION

Smart camera systems powered by high-resolution digital cameras and artificial intelligence (AI) offer unthought opportunities for improving our everyday lives. These new and emerging technologies have a great potential to make public spaces safer, remove friction from shopping experiences or reduce the waiting time in queues.

Applications are endless but commercially available products are still limited. While solutions are available for some applications, like counting customers at store entrances, dwell-time measurement or passenger handling at airports, most everyday activities remain unsupported by intelligent spaces. This is due to the high technical complexity to resolve a certain task, while overall system cost needs to meet ROI requirements. In addition, development of new solutions requires highly specific knowledge in topics like vision system design, embedded software development and machine learning—a combination of experiences which is commonly not available to companies just in need for a vision solution.

As a young technology company, our goal is to innovate in this field in order to provide more affordable solutions to a wider range of applications. By sharing our practical knowledge we hope to support your decision making process and help solving your automation task. This white paper will introduce you to stereo vision, a technology used by many for tasks of monitoring, analyzing and tracking people and their activities. Furthermore, we explain the limitations of conventional stereo vision and present our solution for omnidirectional stereo vision. With this novel technology intelligent video analytics are made available to more complex spaces, while reducing the overall system costs.

TABLE OF CONTENTS

The Rise of Intelligent Video Analytics

Video surveillance or *closed circuit television (CCTV)* has been common for almost a century. Video cameras record non-stop video streams, store them locally or transmit the feed to a control room where dedicated personnel monitors the material. Those systems are very limited in terms of real-time protection of assets and mostly used for forensic purposes.

More recently, *intelligent video analytics (IVA)* allows us to analyze digital video streams in real-time by image understanding software. From automatic intrusion detection in museums and people flow estimation in airports to life-saving emergency detection, this technology opens endless possibilities. These software systems are becoming more and more sophisticated due to technologies like *Deep Neural Networks (DNN)* and *Machine Learning (ML)*.

More advanced cameras provide a spatial understanding of the scene, similar to human vision. Augmenting image data (*2D*) with depth information (*3D*) enables even more powerful *Computer Vision (CV)* algorithms. As an example, object detection becomes more stable due to reliable separation from the background.

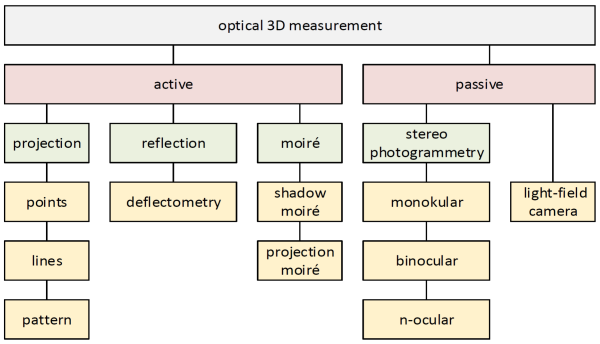

Figure 1: Optical 3D measurement technologies

Spatial sensing can be achieved by various physical principles. Figure 1 provides a brief overview of the most commonly used 3D measurement technologies. However, passive stereo vision has turned out to be the most commonly used method for surveillance applications. This is due to the robustness and simplicity of the technology compared to actively light-emitting counterparts.

Edge Processing for Advanced Privacy Protection

While industry trends were aiming towards cloud solutions for every task of storing or processing data, recent data breaches and privacy issues point into another direction. For analyzing video streams, local (so-called “edge”) processing of image data makes much more sense than cloud computation for two reasons:

1. Sensitive video stored or processed on remote servers is a potential privacy risk for monitored individuals.
2. Large amounts of video data cannot be transferred to remote servers in bandwidth-constraint environments.

The combination of 3D vision, powerful miniaturized processing platforms and advanced image understanding solves both problems. For example, an edge-enabled camera can detect and analyze scenes as well as communicate events like "2 people at position x y" rather than the raw video material.

Limitations of Conventional Rectilinear Cameras

When reviewing state-of-the-art visual surveillance technology and its markets, the system costs turn out to be a crucial issue. Obviously, one of the key concerns is the number of applied cameras for observing a certain scene. Hence, a reduction of necessary sensors (and the attached infrastructure) can have an essential impact on the systems profitability.

A living-room sized space requires up to 2 or 3 conventional cameras for full coverage while avoiding occlusions. Due to their nature of perspective projection, rectilinear cameras are restricted to a *field of view (FOV)* of less than around 120°.

This limitation can be overcome by using omnidirectional cameras, enabled by fish-eye lenses or catadioptric mirrors. While these optical systems introduce a certain distortion, they enable FOVs of more than 180°.

Figure 2: Comparison of a conventional rectilinear camera versus omnidirectional camera in a home environment.

Combining omnidirectional cameras with stereo vision enables a new level of precision in Intelligent Video Analytics. Since accuracy of the most depth sensing principles depends on the distance of an object to the camera,

omnidirectional depth sensing cameras provide another major advantage: If a camera is mounted in the center of a rooms ceiling, the average distance to the scene points is lower than if multiple cameras were mounted in the corners of the room. An example is sketched in Figure 3 where the floor plan of a living environment is shown. The colors indicate the shortest distance (line of sight) to a nearby camera. White areas are covered by neither camera.

As shown, replacing multiple conventional cameras by one

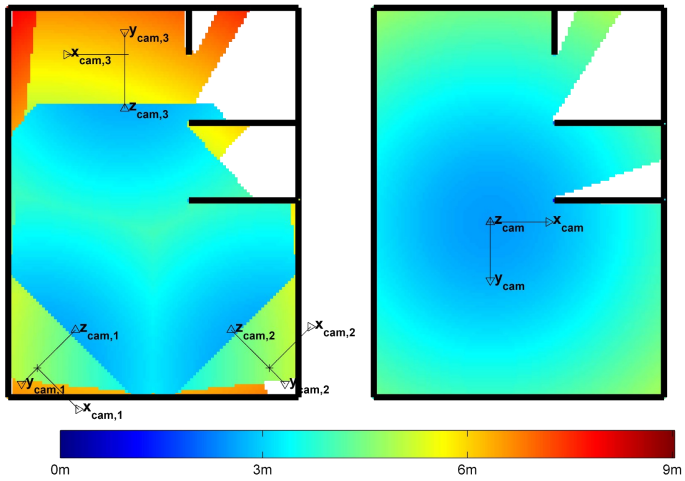

Figure 3: Floor plan of a single-room apartment equipped with optical stereo vision sensors. left: Multiple rectilinear cameras, right: a single

omnidirectional depth sensor reduces the average distance, thereby increasing accuracy and providing homogeneous coverage of the room.

This consideration was the main motivation for the development of HemiStereo®, an omnidirectional depth sensing technology which is able to cover a complete room from floor to ceiling with almost homogeneous measurement accuracy and low computational effort. This technology enables a wide range of applications of intelligent cameras, which we call “Intelligent Spaces”, see Figure 4.

An Introduction to Camera Projection Models

A camera’s projection model describes how an incoming ray of light is projected onto the image sensor. It therefore forms the relation between the 3D world and the 2D image plane of a camera. Most camera systems can be adequately assumed to project radially symmetrically. An incoming light ray is described by its angle of incidence θ (Theta), the angle between the light ray and the camera’s optical axis. This ray is projected onto the image sensor as a point with distance ρ (Rho) from the sensors coordinate origin. Different lens designs employ different mapping functions $\rho(\theta)$, as shown in Figure 5.

Figure 4: Examples of Intelligent Spaces. Top: Automatic detection of emergencies, such as falls of elderly people. Middle: Recognition of fraudulent behavior in security applications. Bottom: Precise in-store analytics for detailed customer insights or automated stores.

Rectilinear cameras follow the perspective projection model. If θ strives towards 90° ($\pi/2$), the imaging function ρ approaches infinity. This leads to peripheral objects being visually extended while central objects are compressed. For this reason, the field of view ($2*\theta_{max}$) of commercially available rectilinear cameras is limited to approximately 120°. However, there are projection models that are not subject to these limitations. These projection functions are compared in Table 1.

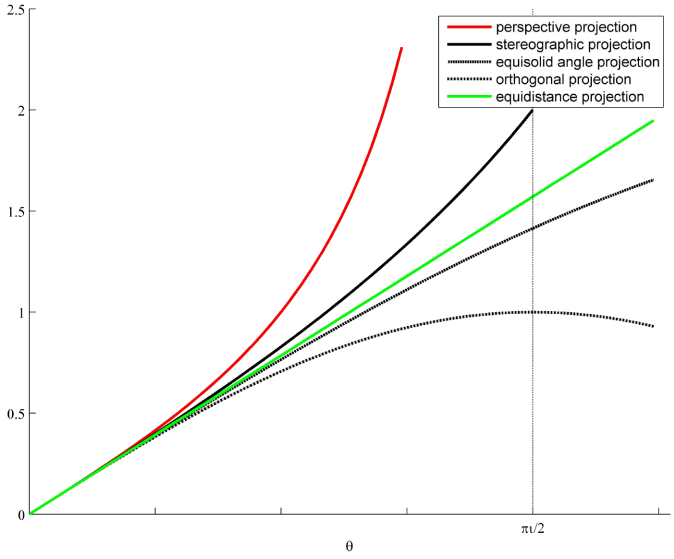

Figure 5: Common camera projection models.

Projection Model	Radial Mapping Function
Perspective (Rectilinear)	$\rho \sim \tan\theta$
Equidistant	$\rho \sim \theta$
Generic	$\rho \sim k_1\theta + k_2\theta^3 + \ldots + k_n\theta^{(2n-1)}$

Table 1: Radial symmetric camera projection models

Equidistant projection describes a linear relationship between θ and ρ. Therefore incoming light rays with an angle of incidence $\theta \geq 90°$ (FOV ≥ 180°) can be mapped onto a finite sensor, thus enabling omnidirectional imaging.

The Science of Conventional Stereo Vision

Stereo vision is a natural concept to derive depth information from two or more views. Conventional stereo setups normally use two imaging devices (*binocular stereo vision*), imitating human eyes. An attached computer processes the signals of the two videos, recreating a spatial perception of the scene, similarly to our brains. The visual displacement of features in the two images correspond to the object distance in 3D. The challenge for a technical implementation is the fast and accurate assignment of corresponding image points, as illustrated in figure 6.

Solving this *correspondence problem* can be very time consuming for a computer. Assuming an image size of 1000 by 1000 pixels (1 megapixel), each single pixel in the left image would have to be compared to 1 million pixels in the right image. This would result in 1 trillion comparison operations per frame.

A straightforward approach to reduce processing time is limiting the search space: Firstly, in the right image every feature was shifted left (by a certain amount depending on the distance). Therefore, the search only needs to be performed left of the pixel position from the left image. Secondly, if the cameras are perfectly aligned, every pixel is shifted only horizontally. Hence the search could be limited to the same scan line. This massively reduces the complexity, because each image pixel in the left image merely has to be compared with a row of pixels in the right image. This method reduces the total number of comparison operations to approximately 1 billion – an acceleration by a factor of 10^3!

Figure 6: The stereo correspondence problem: Finding the matching location in the right image for every pixel in the left image.

While the alignment never could be this accurate in a mechanical assembly of two cameras, it can be reproduced by a process called *image rectification*. This requires two preceding steps: *intrinsic* and *extrinsic calibration*.

Intrinsic calibration: the exact imaging properties of the camera are described using a parameterized mathematical model, containing the previously explained projection model parameters (e.g. k_1 … k_n for the generic projection model, see Table 1). These can be determined automatically with measurement techniques and the use of calibration patterns. Then the image information can be transformed into another camera model using an *image warping*. This compensates for any distortion caused by lenses, which can also be recognized by the pillow effects on the edge of the images in Figure 6.

Extrinsic calibration: The relative position (translation and rotation) of the two cameras to each other is determined by measurement. This can be done with the same calibration patterns used for intrinsic calibration. In the process of rectification, both cameras are virtually rotated in a way that their z-axes (optical axes) are parallel and the x-axes of the camera coordinate system are on one line. Figure 7 shows such a rectified stereo pair whereby the horizontal red line represents a so-called *epipolar line*.

Finally, the displacement information for each pixel in the left image compared to the right is calculated (stereo correspondence). The horizontal offset of corresponding image points $x_{img,l}$ and $x_{img,r}$ is recorded in the image pair as a disparity map (disparity d):

$$d = x_{img,l} - x_{img,r}$$

Solving the correspondence problem is done using *stereo matching*, for which a large number of approaches is available. These methods can be grouped in the following categories:

- *Local stereo matching* compares the neighborhood of a pixel $\boldsymbol{x}_{img,l}$ in the left image to positions $\boldsymbol{x}_{img,r}$ in the right image. Each pixel is processed separately, without taking the full image context into account. This may result in noisy disparity images, especially in nontextured image regions.
- *(Semi-)Global methods* which optimize certain cost functions in the measurement volume x-y-d and thus make certain assumptions for improving the 3D measurement for the nature of the scene. However, the higher precision comes with a much higher computational effort than local methods.
- Stereo matching algorithms based on *deep learning* using *neural networks (CNN)* recently have become more advanced, allowing higher precision than semi-global methods. With the suitable computing architecture (GPU, ASIC) these methods perform more efficiently as well.

If the pixel disparities *d* are calculated for the complete image pair ($\boldsymbol{X}_{img,l}$, $\boldsymbol{X}_{img,r}$), a dense disparity map $\boldsymbol{D}$ for the corresponding color image is obtained as can be seen in Figure 8. Colors indicate the pixels disparity values. Low disparity values (blue) correspond to far points, while high values (green, red) indicate a closer distance to the camera. Black areas in the disparity image indicate areas of image correlation that did not result into a reliable information of pixel displacement.

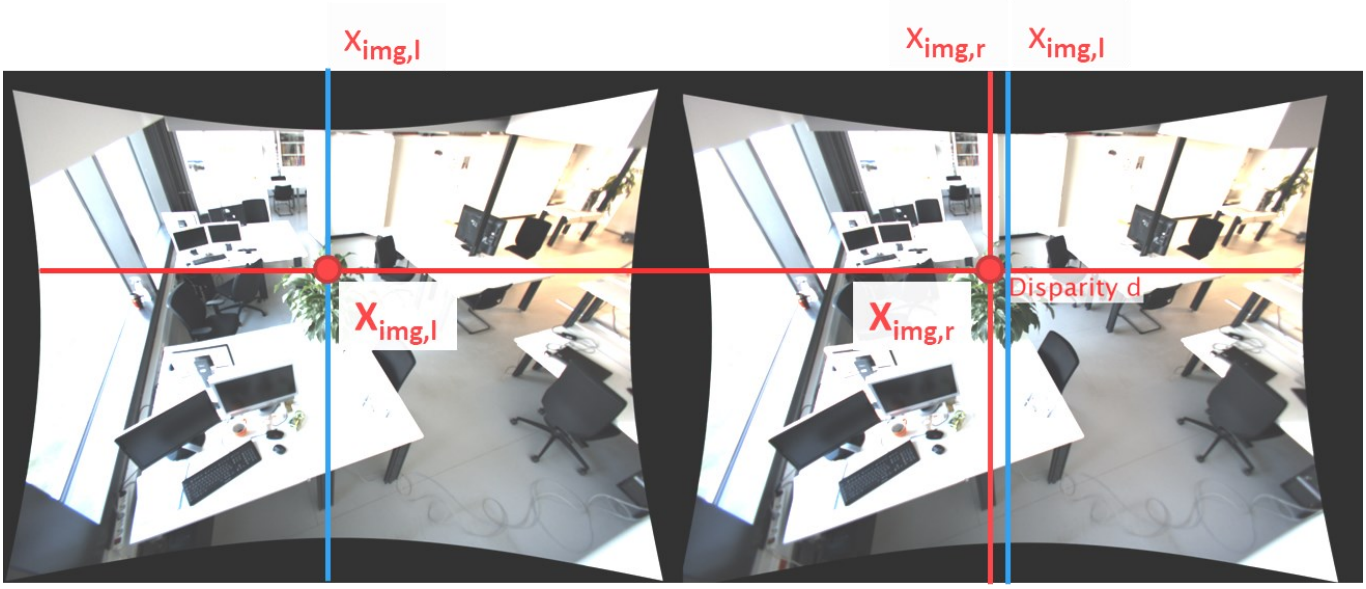

Figure 7: Calculation of pixel disparity based on a rectified image pair.

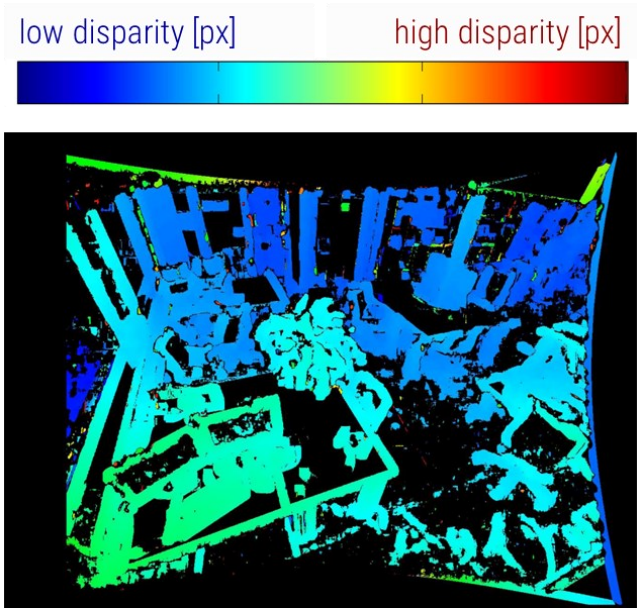

Figure 8: Disparity map ***D***

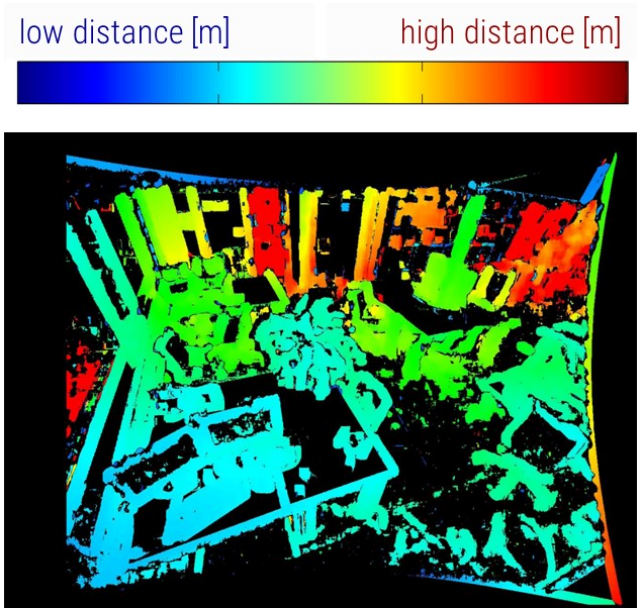

Figure 9: Distance map ***Z***

The metric *distance* (*depth*) value *Z* between each point and the camera can easily be obtained by triangulation. For a rectilinear stereo vision setup the following formula outlines the mathematical relationship between pixel disparity *d* and metric depth information *Z* for each pixel:

$$Z = b \cdot \frac{f}{p_x \cdot d} = b \cdot \frac{\alpha_x}{x_{\text{img,l}} - x_{\text{img,r}}}$$

$$Z_{\min} \Leftrightarrow d_{\max}$$

Further parameters result from the calibration process: the focal length *f*, the physical pixel width p_x and the distance between the two cameras (*base length*) *b*. When transforming each pixel of the disparity map ***D*** using the formula above, we receive the distance map ***Z***, as shown in Figure 9.

Moving to Omnidirectional Stereo Vision

As shown earlier many applications profit from using extreme wide-angle or omnidirectional cameras. Unfortunately these camera systems cannot be used together with perspective (rectilinear) projection models. Fish-eye lenses with an FOV of 180° or more can be described with sufficient accuracy using generic projection models.

However, this introduces significant changes to the beforementioned stereo camera (epipolar) geometry. Epipolar lines are now mapped on so-called *great circles* as shown in Figure 10. This property means that the previously presented stereo matching methods will not work without further preprocessing. Additionally, distance information cannot be calculated for the singularity points.

In order to use fast stereo matching methods, the omnidirectional input image has to be transformed in a way that generates parallel, straight epipolar lines.

A naïve approach would be to extract a *virtual perspective (rectilinear) view* from the omnidirectional image, as shown in Figure 11.

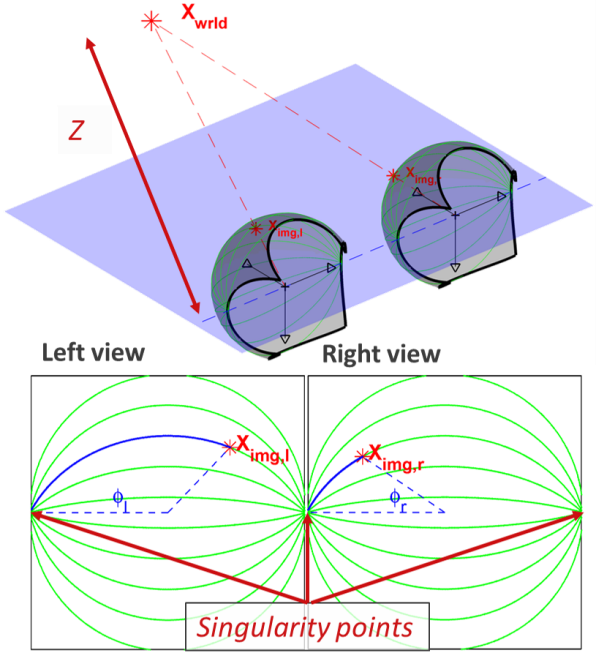

Figure 10: Epipolar geometry in omnidirectional projection. Epipolar lines (i.e. scan lines) are now bended to great circles.

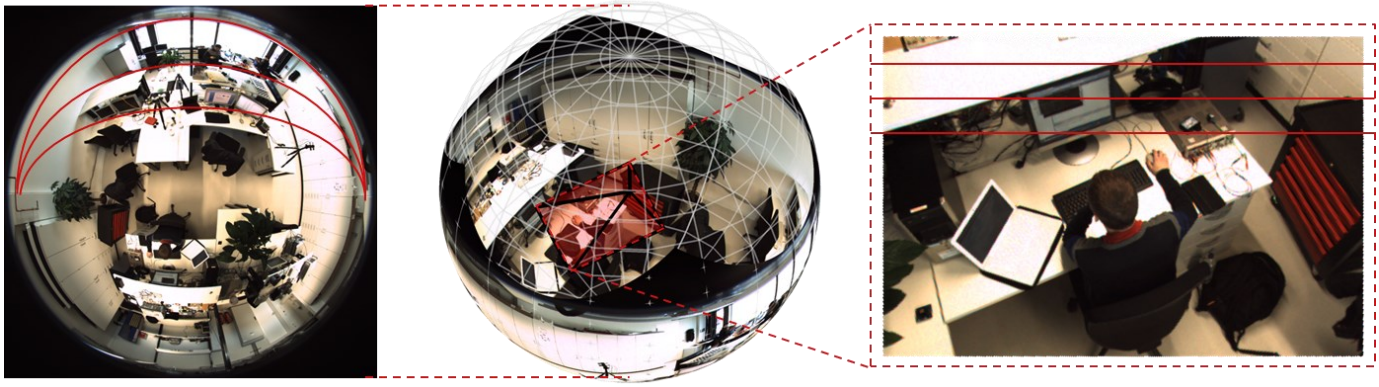

Figure 11: Generating a virtual perspective view from an omnidirectional image through transformation of the mapping model.

While two of those virtual rectilinear views of two cameras can be used for stereo matching, it only allows to compute disparity (depth) information for a certain part of the scene, limited by the rectilinear view.

In order to perform stereo matching for the full omnidirectional image, a specifically tailored projection model is used. The result of this omnidirectional rectification process is shown in Figure 12. Although the noticeable image distortion may look unfamiliar to the human eye, it preserves the desired constraint: straight and parallel epipolar lines. Stereo matching can now take place using two fully rectified images.

Figure 12: Rectified omnidirectional image showing a specific distortion which preserves straight and parallel epipolar lines.

Trinocular Stereo: Full Hemispherical Field-of-View

With a binocular (2-camera) setup the quality of distance information *Z* decreases when approaching the singularity points (*epipoles*), as shown in Figure 13. The smallest measurement error *δZ* can be observed in the center of the omnidirectional image. The error increases continuously along the baseline towards the epipoles. No distance measurement is defined at the singular points and thus the error is infinitely large.

Assuming practical considerations limit the maximum acceptable error to maybe *5*δZ*, the resulting horizonal field of view would be around 155°, as highlighted in Figure 13. While increasing the acceptable maximum error would lead to a higher observable field-of-view, depth measurement for the full hemisphere remains impossible. Further restrictions are caused by

- the maximum disparity parameter, restricting the horizontal FOV;
- strong distortion of the rectified image near the epipoles, prohibiting stereo matching.

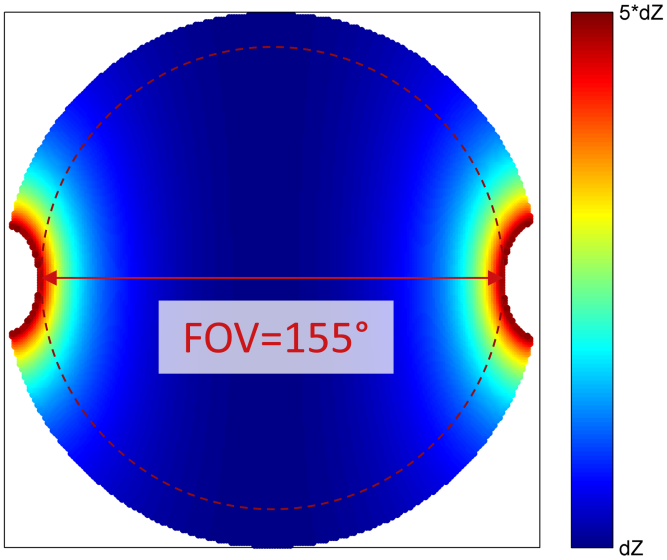

Figure 13: Qualitative error *δZ* of the distance measurement *Z* of a full hemisphere in a binocular stereo setup.

The introduction of an additional camera offers a way out. A *trinocular* (3-camera) stereo vision setup can be constructed from three fish-eye cameras, whereby all cameras should be mounted on one level and the optical axes should be aligned in parallel. Using this configuration HemiStereo® DK1 is able to provide a fully hemispherical field-of-view of 180° x 180° (HxV).

Two cameras each form a binocular stereo camera and cover a small part of the hemisphere to be measured. By merging all partial distance maps we achieve full hemispherical distance information. This leads to a highly accurate hemispherical depth map.

All these computations are already performed in real-time inside HemiStereo® DK1 using a powerful NVIDIA® Jetson™ TX2 System-on-Module. Users can monitor this data through the remote viewer app as shown in Figure 15 or develop custom IVA applications using omnidirectional depth sensing and *Artificial Intelligence (AI)*.

Figure 14: Trinocular stereo vision setup used in the HemiStereo® DK1 depth sensing camera.

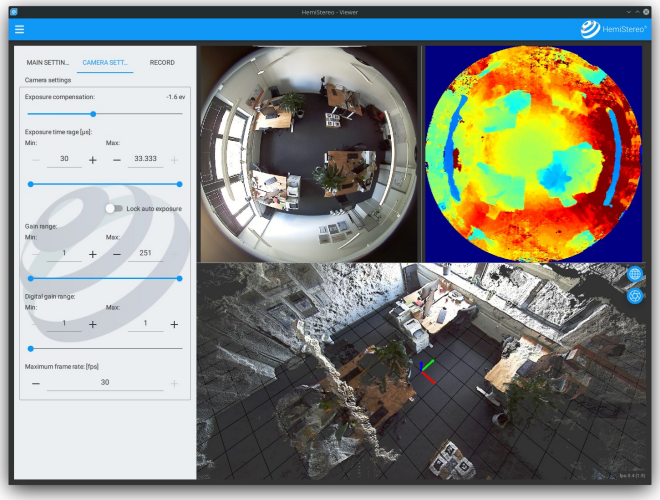

Figure 15: Remote monitoring of omnidirectional images, depth map and reconstructed 3D point clouds.

Get your solution now!

You want to learn more about depth sensing, vision system design or intelligent video analytics? You're ready to start building your own intelligent spaces application now or you need additional guidance?

We are happy to assist you in realizing your own project. 3dvisionlabs offers a variety of different depth sensing products as well as vision system design services.

Head over to our website or talk to one of our vision designers or AI experts today:

Web: 3dvisionlabs.com

Mail: solutions@3dvisionlabs.com

Phone: +49 (0) 371 3371 6555

ABOUT 3DVISIONLABS

We are a young technology start-up company based in Chemnitz, Germany. At 3dvisionlabs we love what we do: We develop the next generation of depth camera technologies. With our innovative products we give future intelligent environments and robots a superhuman sense of vision. With HemiStereo® we enable new applications of AI-powered perception in areas like Smart Buildings, Robot Navigation or Retail Automation.