Reginald L. Lagendijk, et. Al. “Stereoscopic Image Processing.”
2000 CRC Press LLC. <http://www.engnetbase.com>.

Stereoscopic Image Processing[1]

Reginald L. Lagendijk
Delft University of Technology

Ruggero E.H. Franich
AEA Technology,
Culham Laboratory

Emile A. Hendriks
Delft University of Technology

57.1 Introduction

Static images and dynamic image sequences are the projection of time-varying three-dimensional real world scenes onto a two-dimensional plane. As a result of this planar projection, depth information of objects in the scene is generally lost. Only by *cues* such as shadow, relative size and sharpness, interposition, perspective factors, and object motion, can we form an impression of the depth organization of the real world scene.

In a wide variety of image processing applications, explicit depth information is required in addition to the scene's gray value information (representing intensities, color, densities, etc.) [2, 4, 7]. Examples of such applications are found in 3-D vision (robot vision, photogrammetry, remote sensing systems); in medical imaging (computer tomography, magnetic resonance imaging, microsurgery); in remote handling of objects, for instance in inaccessible industrial plants or in space exploration; and in visual communications aiming at virtual presence (conferencing, education, virtual travel and shopping, virtual reality). In each of these cases, depth information is essential for accurate image analysis or for enhancing the realism. In remote sensing the terrain's elevation needs to be accurately determined for map production, in remote handling an operator needs to have precise knowledge of the three-dimensional organization of the area to avoid collisions and misplacements, and in visual communications the quality and ease of information exchange significantly benefits from the high degree of realism provided by scenes with depth.

Depth in real world scenes can be explicitly measured by a number of range sensing devices such as by laser range sensors, structured light, or ultrasound. Often it is, however, undesirable or unnecessary to have separate systems for acquiring the intensity and the depth information because

[1]This work was supported in part by the European Union under the RACE-II project DISTIMA and the ACTS project PANORAMA.

of the relative low resolution of the range sensing devices and because of the question of how to fuse information from different types of sensors.

An often used alternative to acquire depth information is to record the real world scene from different perspective viewpoints. In this way, multiple images or (preferably time-synchronized) image sequences are obtained that implicitly contain the scene's depth information. In the case that multiple views of a single scene are taken without any specific relation between the spatial positions of the viewpoints, such recordings are called *multiview images*. Generally speaking, when recordings are obtained from an increasing number of different viewpoints, the 3-D surfaces and/or interior structures of the real world scene can be reconstructed more accurately. The terms *stereoscopic image* and *stereoscopic image sequence* are reserved for the special case that two perspective viewpoints are recorded or computed such that they can be viewed by a human observer to produce the effect of natural depth perception (see Fig. 57.1). Therefore, the two views are required to be recorded under specific constraints such as the cameras' separation, convergence angle, and alignment [8]. Stereoscopic images are not truly 3-D images since they merely contain information about the 2-D projected real world surfaces plus the depth information at the perspective viewpoints. They are, therefore, sometimes called 2.5-D images.

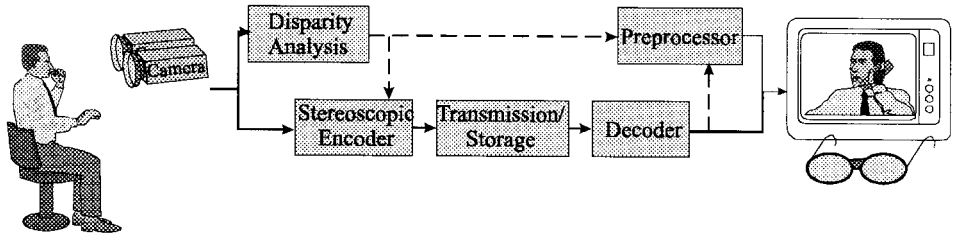


FIGURE 57.1: Illustration of system for stereoscopic image (sequence) recording, processing, transmission, and display.

In the broadest meaning of the word, a digital stereoscopic system contains the following components: stereoscopic camera setup, depth analysis of the digitized and recorded views, compression, transmission or storage, decompression, preprocessing prior to display, and, finally, the stereoscopic display system. The emphasis here is on the image processing components of this stereoscopic system; that is, depth analysis, compression, and preprocessing prior to the stereoscopic display. Nonetheless, we first briefly review the perceptual basis for stereoscopic systems and techniques for stereoscopic recording and display in Section 57.2. The issue of depth or *disparity analysis* of stereoscopic images is discussed in Section 57.3, followed by the application of compression techniques to stereoscopic images in Section 57.4. Finally, Section 57.5 considers the issue of stereoscopic image interpolation as a preprocessing step required for multiviewpoint stereoscopic display systems.

57.2 Acquisition and Display of Stereoscopic Images

The human perception of depth is brought about by the hardly understood brain process of fusing two planar images obtained from slightly different perspective viewpoints. Due to the different viewpoint of each eye, a small horizontal shift exists, called *disparity*, between corresponding image points in the left and right view images on the retinas. In stereoscopic vision, the objects to which the eyes are focused and accommodated have zero disparity, while objects to the front and to the back have negative and positive disparity, respectively, as is illustrated in Figure 57.2. The differences in

disparity are interpreted by the brain as differences in depth ΔZ.

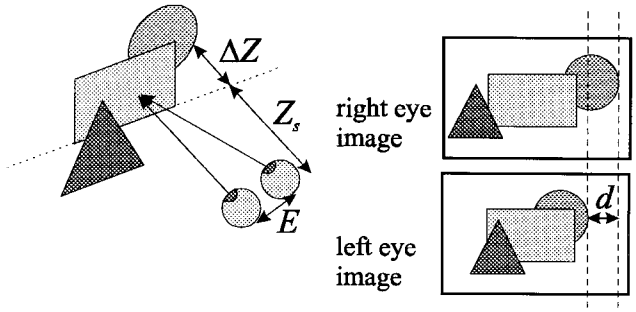


FIGURE 57.2: Stereoscopic vision, resulting in different disparities depending on depth.

In order to be able to perceive depth using recorded images, a stereoscopic camera is required which consists of two cameras that capture two different, horizontally shifted perspective viewpoints. This results in a shift (or disparity) of objects in the recorded scene between the left and the right view depending on their depth. In most cases, the interaxial separation or baseline B between the two lenses of the stereoscopic camera is in the same order as the eye distance E (6 to 8 cm). In a simple camera model, the optical axes are assumed to be parallel. The depth Z and disparity d are then related as follows:

$$d = \lambda \frac{B}{\lambda - Z} , \tag{57.1}$$

where λ is the focal length of the cameras. Fig. 57.3(a) illustrates this relation for a camera with $B = 0.1$ m and $\lambda = 0.05$ m. A more complicated camera model takes into account the convergence of the camera axes with angle β. The resulting relation between depth and disparity, which is a much more elaborate expression in this case, is illustrated in Fig. 57.3(b) for the same camera parameters and $\beta = 1°$. It shows that, in this case, the disparity is not only dependent on the depth Z of an object, but also on the horizontal object position X. Furthermore, a converging camera configuration also leads to small vertical disparity components, which are, however, often ignored in subsequent processing of the stereoscopic data. Figures. 57.4(a) and (b) show as an example a pair of stereoscopic images encountered in video communications.

When recording stereoscopic image sequences, the camera setup should be such that, when displaying the stereoscopic images, the resulting shifts between corresponding points in the left and right view images on the display screen allow for comfortable viewing. If the observer is at a distance Z_s from the screen, then the observed depth Z_{obs} and displayed disparity d are related as:

$$Z_{\text{obs}} = Z_s \frac{E}{E - d} . \tag{57.2}$$

In the case that the camera position and focusing are changing dynamically, as is the case, for instance, in stereoscopic television production where the stereoscopic camera may be zooming, the camera geometry is controlled by a set of production rules. If the recorded images are to be used for multiviewpoint stereoscopic display, a larger interaxial lens separation needs to be used, sometimes even up to 1 m. In any case, the camera setup should be geometrically calibrated such that the two cameras capture the same part of the real world scene. Furthermore, the two cameras and A/D converters need to be electronically calibrated to avoid unbalances in gray value of corresponding points in the left and right view image.

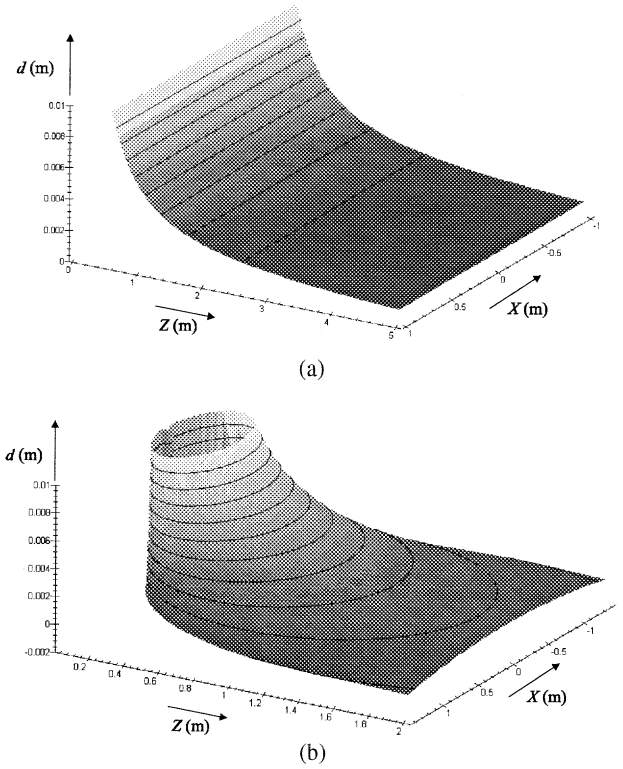


FIGURE 57.3: (a) Disparity as a function of depth for a sample parallel camera configuration; (b) disparity for a sample converging camera configuration.

The stereoscopic image pair should be presented such that each perspective viewpoint is seen only by one of the eyes. Most practical state-of-the-art systems require viewers to wear special viewing glasses [6]. In a *time-parallel* display system, the left and right view images are presented simultaneously to the viewer. The views are separated by passive viewing glasses such as red-green viewing glasses requiring the left and right view to be displayed in red and green, respectively, or polarized viewing glasses requiring different polarization of the two views. In a *time-sequential* stereoscopic display, the left and right view images are multiplexed in time and displayed at a double field rate, for instance 100 or 120 Hz. The views are separated by means of the active synchronized shuttered glasses that open and close the left and right eyeglasses depending on the viewpoint being shown. Alternatively, lenticular display screens can be used to create spatial interference patterns such that the left and right view images are projected directly into the viewer's eyes. This avoids the need of wearing viewing glasses.

57.3 Disparity Estimation

The key difference between planar and stereoscopic images and image sequences is that the latter implicitly contains depth information in the form of disparity between the left and right view images. Not only is the presence of disparity information essential to the ability of humans to perceive depth, disparity can also be exploited for automated depth segmentation of real world scenes, and for compression and interpolation of stereoscopic images or image sequences [1].

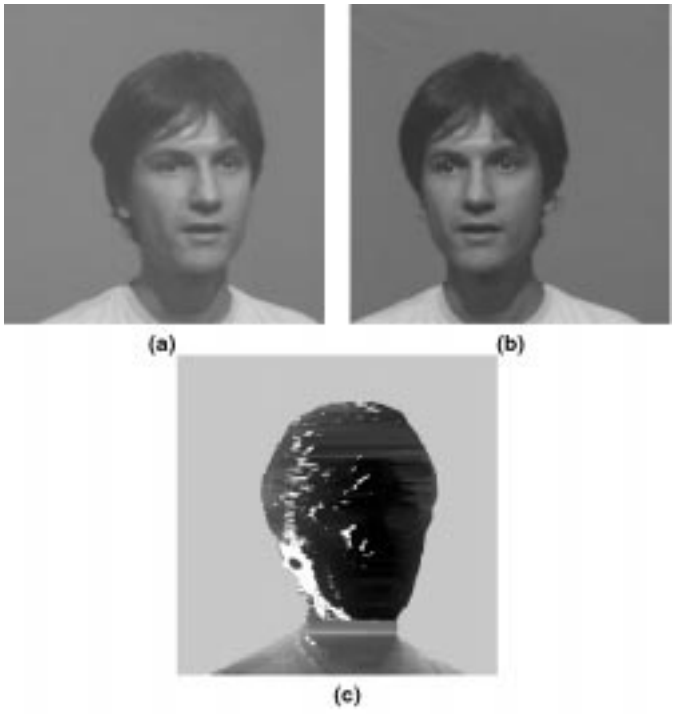


FIGURE 57.4: The left (a) and right (b) view image from a stereoscopic image pair. (c) Disparity field in the stereoscopic image pair represented as gray values (black is foreground, gray is background, white is occlusion).

To be able to exploit disparity information in a stereoscopic pair in image processing applications, the relation between the contents of the left view image and the right view image has to be established, yielding the *disparity (vector) field*. The disparity field indicates for each point in the left view image the relative shift of the corresponding point in the right view image and vice versa. Since some parts of one view image may not be visible in the alternate view image due to occlusion, not all points in the image pair can be assigned a disparity vector.

Disparity estimation is essentially a correspondence problem. The correspondence between the two images can be determined by either matching features or by operating on or matching of small patches of gray values. Feature matching requires as a preprocessing step the extraction of appropriate features from the images, such as object edges and corners. After obtaining the features, the correspondence problem is first solved for the spatial locations at which the features occur, from which next the full disparity field can be deduced by, for instance, interpolation or segmentation procedures. Feature-based disparity estimation is especially useful in the analysis of scenes for robot vision applications [4, 11].

Disparity field estimation by operating directly on the image gray value information is not unlike the problem of motion estimation [11, 12]. The first difference is that disparity vectors are approximately horizontally oriented. Deviations from the horizontal orientation are caused by the convergence of the camera axes and by differences between the camera optics. Usually vertical disparity components are either ignored or rectified. A second difference is that disparity vectors can take on a much larger range of values within a single image pair. Furthermore, the disparity field may have large discontinuities associated with objects neighboring in the planar projection but having a very much different depth. In those regions of the stereoscopic image pair where one finds large discontinuities in the disparity

field due to abrupt depth changes, large regions of occlusion will be present. Estimation methods for disparity fields must therefore be able not only to find the correspondence between information in the left and right view images, but must also be able to detect and handle discontinuities and occlusions [1].

Most disparity estimation algorithms used in stereoscopic communications rely on matching small patches of gray values from one view to the gray values in the alternate view. The matching of this small patch is not carried out in the entire alternate image, but only within a relatively small search region to limit the computational complexity. Standard methods typically use a rectangular match block of relatively small size (e.g., 8×8 pixels), as illustrated in Fig. 57.5. The relative horizontal

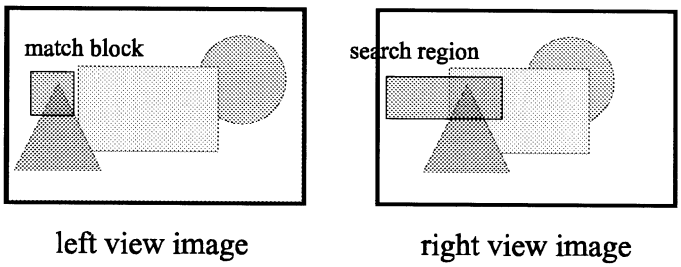


FIGURE 57.5: Block matching disparity estimation procedure by comparing a match block from the left image to the blocks within a horizontally oriented search region in the right image.

shift between a match block and the block within the search region of the alternate image that results in the smallest value of a criterion function used is then assigned as disparity vector to the center of that match block. Often used criterion functions are the sum of squares and the sum of the absolutes values of the differences between the gray values in the match block and the block being considered in the search region [3, 12].

The above procedure is carried out for all pixels, first matching the blocks from the left view image to the right view image, then vice versa. From the combination of the two resulting disparity fields and the values of the criterion function, the final disparity field is computed, and occluding areas in the stereoscopic image pair are detected. For instance, one way of detecting occlusions is a local abrupt increase of the criterion function, indicating that no acceptable correspondence between the two image pairs could be found locally. Fig. 57.4(c) illustrates the result of a disparity estimation process as an image in which different gray values correspond to different disparities (and thus depth), and in which "white" indicates occluding regions that can be seen in the left view image but that cannot be seen in the right view image.

More advanced versions of the above *block matching disparity estimator* use hierarchical or recursive approaches to improve the consistency or smoothness of the resulting disparity field, or are based on the optical flow model often used in motion estimation. Other approaches use preprocessing steps to determine the dominant disparity values that are then used as candidate solutions during the actual estimation procedure. Finally, most recent approaches use advanced Markov random field models for the disparity field and/or they make use of more complicated cost functions such as the *disparity space image*. These approaches typically require exhaustive optimization procedures but they have the potential of accurately estimating large discontinuities and of precisely detecting the presence of occluding regions [1].

In image analysis problems, disparity estimation is often considered in combination with the segmentation of the stereoscopic image pair. Joint disparity estimation and texture segmentation

methods partition the image pair into spatially homogeneous regions of approximately equal depth. Disparity estimation in image sequences is typically carried out independently on successive frame pairs. Nevertheless, the need for temporal consistency of successive disparity fields often requires temporal dependencies to be exploited by postprocessing of the disparity fields. If an image sequence is recorded as an interlaced video signal, disparity estimation should be carried out on the individual fields instead of frames to avoid confusion between motion displacements and disparity.

57.4 Compression of Stereoscopic Images

Compression of digital images and image sequences is necessary to limit the required transmission bandwidth or storage capacity [3, 5]. One of the compression principles underlying the JPEG and MPEG standards is to avoid transmitting or storing gray value information that is predictable from the signal's spatial or temporal past, i.e., information that is redundant. In both JPEG and MPEG, this principle is exploited by a spatial DPCM system, while in MPEG motion-compensated temporal prediction is also used to exploit temporal redundancies.

When dealing with stereoscopic image pairs, a third dimension of redundancy appears, namely the mutual predictability of the two perspective views [9]. Although the left and right view images are not identical, gray value information in, for instance, the left view image is highly predictable from the right view image if the horizontal shift of corresponding points, i.e., the disparity, is taken into account. Thus, instead of transmitting or storing both views of a stereoscopic image pair, only the right view image is retained, together with the disparity field. Since the construction of the left view image from the right view is not perfect due to errors in the estimated disparity field and due to presence of occluding areas and perspective differences, some information of the *disparity-compensated prediction error* of the left view (i.e., the difference between the predicted gray values and the actual gray values in the left view image) also needs to be retained. Figure 57.6 shows the *disparity-compensated prediction* and the *disparity-compensated prediction error* of the left view image from Fig. 57.4(a) using the right view image in Fig. 57.4(b) and the disparity field in Fig. 57.4(c). In most cases, the sum of the bit rates needed for coding the disparity vector field and the disparity-compensated prediction error is much smaller than the bit rate needed for the left view image when compressed without disparity compensation.

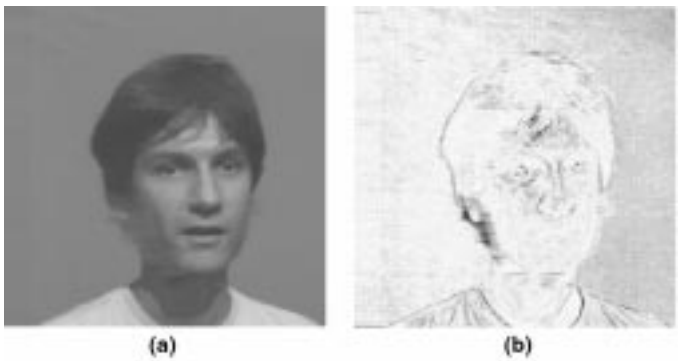


FIGURE 57.6: (a) Disparity-compensated prediction and (b) disparity-compensated prediction error of the left view image (scaled for maximal visibility) in Fig. 57.4. Black areas indicate a large error.

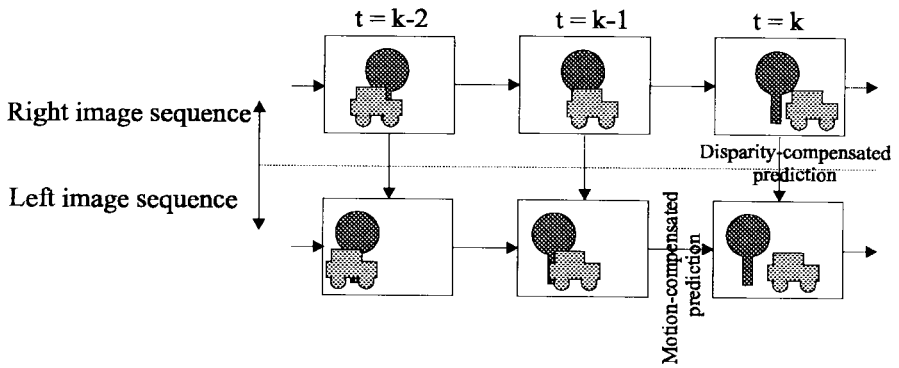
In image sequence, left view images can be compressed efficiently by carrying out motion-compensated prediction from previous left view images, by disparity-compensated prediction from the corresponding right view image, or by a combination of the two by choosing for motion-compensation or disparity-compensation on a block-by-block basis, as illustrated in Fig. 57.7(a). Basically this is a direct extension of the MPEG compression standard with an additional prediction mode for the left view image sequence. The effect of this additional (disparity-compensated) prediction mode is that the variance of the prediction error of the left view image sequence is further decreased [see Fig. 57.7(b)], meaning that more compression of the left view sequence is possible than when independently compressing the two views of the stereoscopic sequence. Figure 57.8 schematically shows the architecture of a disparity- and motion-compensated encoder for stereoscopic video.

57.5 Intermediate Viewpoint Interpolation

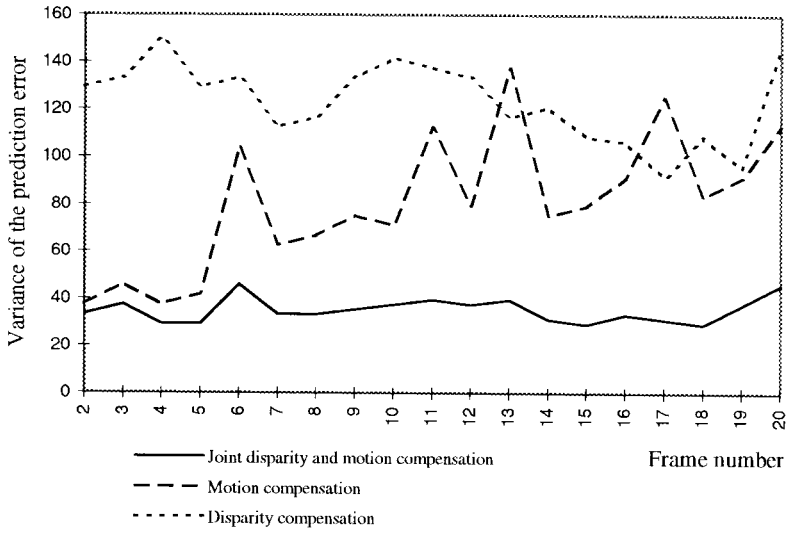
The system illustrated in Fig. 57.1 assumes that the stereoscopic image captured by the cameras is directly displayed at the receiver's end. One of the shortcomings of such a two-channel stereoscopic system is that shape and depth distortion occur when the stereoscopic images are viewed from an off-center position. Furthermore, since the cameras are in a fixed position, the viewer's (horizontal) movements do not provide additional information about, for instance, objects that are partly occluded. The lack of this "look around" capability especially is a limiting factor in the truly realistic visualization of a recorded real world scene.

In a multi-channel or *multiview stereoscopic system*, multiple viewpoints of the same real world scene are available. The stereoscopic display then shows only those two perspective views which correspond as well as possible with the viewer's position. To this end some form of tracking the viewer's position is necessary. The additional viewpoints could be obtained by installing more cameras at a wide range of possible viewpoints. On grounds of complexity and costs the number of cameras will typically be limited to three to five, meaning that not all possible positions of the viewer are covered in this way. If, because of the viewer's position, a view of the scene is needed from an unavailable camera position, a *virtual camera* or *intermediate viewpoint* must be constructed from the available camera viewpoints (see Fig. 57.9).

The construction of intermediate viewpoints is an interpolation problem, which has much in common with the problem of video standards conversion [11]. In its most simple form, the interpolated viewpoint is merely a weighted average between the images from the nearest two camera viewpoints, which are called the key images. Such a straightforward averaging ignores the presence of disparity between the key images, yielding a highly blurred and essentially useless result [see Fig. 57.10(a)]. If, however, the disparity vector field between the two key images has been estimated and the areas of occlusions are known, the interpolation can be carried out along the *disparity axis*, such that the disparity information in the interpolated image corresponds exactly to the virtual camera position. For the points where a correspondence exists between the two key images, this construction process is called *disparity-compensated interpolation*, while for the occluding regions *extrapolation* has to be carried out from the key images [10]. Figure 57.10(b) illustrates the result of intermediate viewpoint interpolation on the stereoscopic image pair in Figs. 57.4(a) and (b).



(a)



(b)

FIGURE 57.7: (a) Principle of joint disparity- and motion-compensated prediction for the left view of a stereoscopic image sequence; (b) variance of the prediction error of the left view image sequence when using motion-compensation, disparity-compensation, or joint motion-disparity compensation on a block-by-block basis.

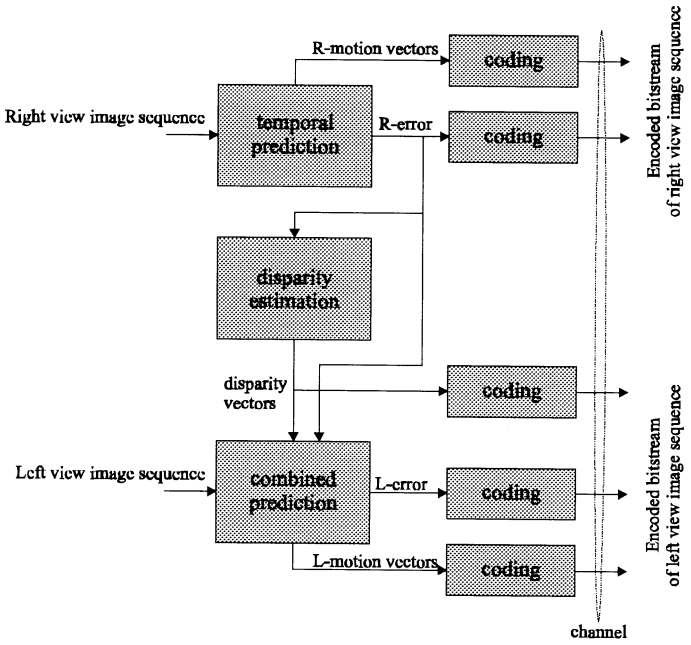


FIGURE 57.8: Architecture of a disparity- and motion-compensated encoder for stereoscopic video.

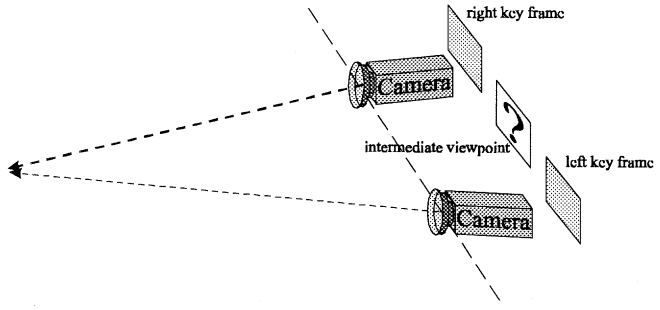


FIGURE 57.9: Multiview stereoscopic system with interpolated intermediate viewpoint (virtual camera).

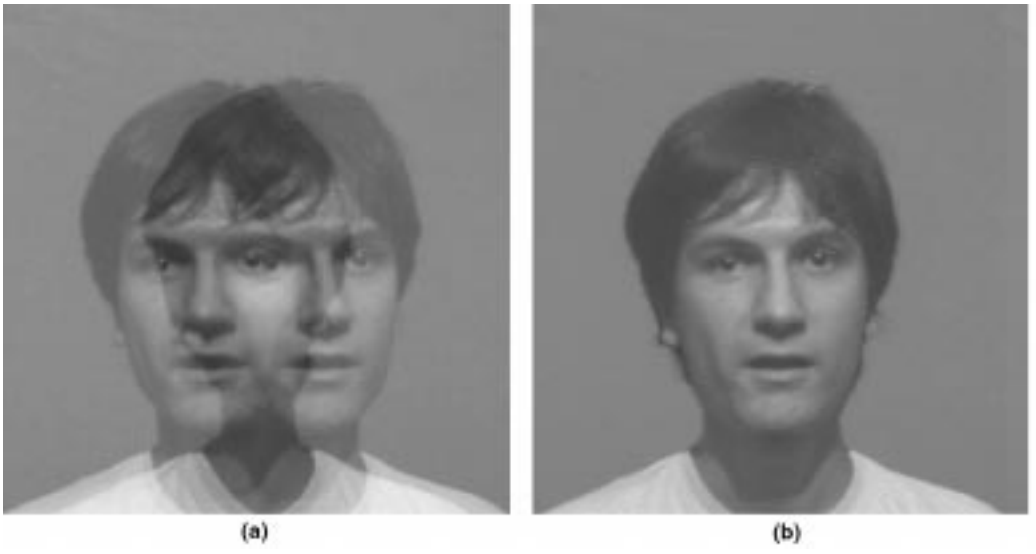


FIGURE 57.10: Interpolation of an intermediate viewpoint image of the stereoscopic pair in Fig. 57.4: (a) without and (b) with taking into account the disparity information between the key frames.

References

[1] *Proceedings of the 1995 International Workshop on Stereoscopic and Three Dimensional Imaging*, Efstratiadis, S. et al., Eds., Santorini, Greece, 1995.

[2] Dhond, U.R. and Aggerwal, J.K., Structure from stereo, *IEEE Trans. on System, Man and Cybernetics*, 19(6), 1489-1509, 1989.

[3] Hang, H.-M and Woods, J.W., *Handbook of Visual Communications*, Academic Press, San Diego, CA, 1995.

[4] Horn, B.K.P., *Robot Vision*, MIT Press, Cambridge, 1986.

[5] Jayant, N.S. and Noll, P., *Digital Coding of Waveforms*, Prentice-Hall, London, 1984.

[6] Lipton, L., *The Crystal Eyes Handbook*, StereoGraphics Corporation, 1991.

[7] Marr, D., *Vision*, Freeman, San Francisco, 1982.

[8] Pastoor, S., 3-D television: A survey of recent research results on subjective requirements, *Signal Processing: Image Communications*, 4(1), 21-32, 1991.

[9] Perkins, M.G., Data compression of stereopairs, *IEEE Trans. Commun.*, 40(4), 684-696, 1992.

[10] Skerjanc, R. and Liu, J., A three camera approach for calculating disparity and synthesizing intermediate pictures, *Signal Processing: Image Communications*, 4(1), 55-64, 1991.

[11] Tekalp, A.M., *Digital Video Processing*, Prentice-Hall, Upper Saddle River, NJ, 1995.

[12] Tziritas, G. and Labit, C., *Motion Analysis for Image Sequence Coding*, Elsevier, Amsterdam, 1994.