Interactive 3-DTV—Concepts and Key Technologies

CHRISTOPH FEHN, RENÉ DE LA BARRÉ, AND SIEGMUND PASTOOR

Invited Paper

The objective of this paper is to provide an overview about recent trends in the area of three-dimensional television (3-DTV). This includes the application of new 3-D data representation formats, which are inherently interactive and more flexible than the traditional (two-view) stereoscopic image. In this context, we describe an experimental 3-DTV system that is based on the joint distribution of monoscopic color video and associated per-pixel depth information. From these data, one or more "virtual" views of a real-world scene can be synthesized in real-time at the receiver side (i.e., in a 3-DTV set-top box) by means of so-called depth-image-based rendering (DIBR) techniques. In addition, the paper provides details on the latest advances in glasses-free (autostereoscopic) 3-DTV display development, both for single and multiple users, as well as on multimodel user interfaces based on head, gaze, or gesture tracking.

Keywords*—(Auto)stereoscopic 3-D displays, coding and broadcast transmission, depth-image-based rendering (DIBR), interactive media, three-dimensional television (3-DTV).*

I. INTRODUCTION

Three-dimensional television (3-DTV) is believed by many to be the next logical development toward a more natural and life-like visual home entertainment experience. Although the basic technical principles of stereoscopic TV were already demonstrated in the 1920s by John Logie Baird [1], the step into the third dimension still remains to be taken. The many different reasons that have hampered a successful introduction of 3-DTV so far could now be overcome by recent advances in a number of key technologies, with the following developments being of particular importance:

1) the introduction and increasing propagation of digital broadcast TV in Europe, Asia, and the United States [2];
2) the convergence of the two former separate worlds of television and PC [3], [4];
3) the growing availability of ubiquitious broad-band IP networks as powerful new distribution channels;
4) the recent progress in the area of efficient image-based modeling (IBM) and rendering (IBR) techniques [5];
5) the promising latest achievements in the area of single and multiple-user autostereoscopic 3-D display technologies [6]–[8];
6) the increased interest in human factors requirements for high-quality 3-DTV systems [9].

Since 1838, the year Sir Charles Wheatstone invented the mirror stereoscope, it is known that compelling 3-D reproductions of real-world scenes can be created by providing a user with two separate images—one for the left eye and one for the right eye—which have been captured from slightly different viewing positions [10]. Today, binocular stereopsis is still appreciated as the most impressive way of perceiving 3-D scenes; however, the classic two-view image format is no longer regarded to be the best way of representing 3-D spatial information. In fact, a range of different 3-D data representations have been developed over the past years providing greater flexibility and a potential for interactivity that had not been thinkable with conventional stereoscopic imagery [11]. (A detailed overview and categorization of state-of-the-art 3-D scene representation formats can be found in the seminal work of Shum and Kang [5].)

Texture-mapped 3-D meshes, for example, which are known from classical 3-D computer graphics [12], are very well suited to approximate the geometry and surface properties of (usually static) human-made objects for interactive visual exploration in either 2-D or 3-D. Volumetric scene representations such as voxels (short for volume element), in turn, can be used to efficiently model dynamic 3-D video objects, which can then be examined from virtually any required viewing direction [13]. Similarly, the enhancement of one or more color video sequences with per-pixel depth information can permit the viewer to synthesize and display "virtual" stereoscopic images customized for his/her 3-D display device and/or viewing preferences [14]. Even on a

Manuscript received February 1, 2005; revised June 28, 2005.

The authors are with the Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institut (HHI), Berlin 10587, Germany (e-mail: fehn@hhi.fhg.de; barre@hhi.fhg.de; pastoor@hhi.fhg.de).

Digital Object Identifier 10.1109/JPROC.2006.870688

regular 2-D screen, the same 3-D data format can be utilized to provide head-motion parallax support for interactive Matrix-like 3-D viewing effects [15].

While all the 3-D data representations discussed so far depend on detailed knowledge of the scene geometry, simply increasing the number of views can also lead to interesting new applications. For instance, multiview images, captured from a number of closely spaced viewpoints (e.g., 9–16), can serve as input for autostereoscopic 3-DTV displays for multiple viewers [8]. Providing even more images (in the dozens or even hundreds) actually facilitates the synthesis of high-quality intermediate "virtual" views without the need to acquire explicit 3-D structure [16]. Although the vast complexity in terms of data acquisition, storage, and transmission currently limits the feasibility of such *light fields* for most interactive 3-D video applications, they can still be very useful for the 3-D reproduction of small, static objects with highly complex depth characteristics (such as plants).

Some of these 3-D technologies are already widely used in commercial applications (e.g., texture-mapped 3-D meshes in 3-D gaming, computer-aided design (CAD), and interactive product advertisements or volumetric models in medical imaging and visualization). On the other hand, their impact on our television experiences has been negligible until today. In the future, however, this situation might change, especially due to the impact of two recent initiatives.

1) Initiated by the five major Japanese electronics companies—Itochu, NTT Data, Sanyo Electric, Sharp, and Sony—a 3-D consortium was founded in March 2003 with the goal to enhance the potential market for 3-D technologies [17]. Subcommittees within this consortium, which at present unites over 200 members worldwide, discuss issues such as spreading image formats appropriate for various applications and I/O devices as well as developing guidelines and authoring tools for 3-D content creation.
2) Furthermore, the MPEG committee established a new *ad hoc* group on 3-D audio/visual coding (3DAV) [11]. This activity, started in December 2001, aims at the development of improved compression technologies for novel 3-D applications such as omnidirectional video, free viewpoint television (FTV) based on light fields, and 3-DTV using novel depth-image-based rendering (DIBR) methods.

This paper is organized in five distinct parts. Following this introduction, Section II describes the signal processing and data transmission chain of an experimental 3-DTV system that is based on a 3-D data representation format consisting of monoscopic color video and associated per-pixel depth information. This paper addresses all aspects from the creation of the 3-D content to coding and transmission as well as the efficient synthesis of "virtual" stereoscopic views. This is followed in Section III with a detailed overview of the current state of the art in (glasses-free) autostereoscopic 3-DTV display technology. Thereafter, Section IV looks at advanced input devices, which might allow future 3-DTV viewers to interact more naturally (e.g., through speech, gaze, or gesturing) with the provided 3-D information. Finally, Section V concludes the paper and provides an outlook on further developments in this exciting field.

Fig. 1. 3-D data representation format consisting of: (a) regular 2-D color video in digital TV format (e.g., in Europe 720 × 576 luminance pels, 25 Hz, interlaced); (b) accompanying 8-bit depth-images with the same spatio-temporal resolution. The depth-images are normalized to a near clipping plane Z_{near} and a far clipping plane Z_{far}.

II. Image Capture, Signal Processing, and Transmission

This section describes a new, experimental 3-DTV system that has been developed as part of the European IST project Advanced Three-Dimensional Television System Technologies (ATTEST) [18]. In contrast to former proposals, which in most cases relied on the basic concept of an end-to-end stereoscopic video chain (see, e.g., [19], [20]), this approach is based on a more flexible distribution of an image-based layered 3-D data representation format consisting of monoscopic color video and associated per-pixel depth information (Fig. 1).

Each of these *depth-images* stores depth information as 8-bit grayvalues with the graylevel 0 specifying the furthest value and the graylevel 255 defining the closest value. To translate this data format to real, metric depth values—which are required for the "virtual" view generation (cf. Section II-B)—and to be flexible with respect to 3-D scenes with different depth characteristics, the grayvalues are specified in reference to two main depth clipping planes. The *near clipping plane* Z_{near} (graylevel 255) defines the smallest metric depth value Z that can be represented in the particular depth-image. Accordingly, the *far clipping plane* Z_{far} (graylevel 0) defines the largest representable metric depth value. In case of a uniform quantization of depth, all $N = 2^8$ discrete *depth layers* can be calculated from these two extremes as

$$Z_\nu = Z_{\mathrm{near}}\left(\frac{\nu}{N-1}\right) + Z_{\mathrm{far}}\left(1 - \frac{\nu}{N-1}\right) \qquad (1)$$

where $\nu \in [0, \ldots, N-1]$ specifies the respective grayvalue.

From this 3-D data representation, one or more "virtual" views of a real-world 3-D scene can then be generated in real-time at the *receiver side* by means of so-called DIBR techniques. To ensure a backward-compatible distribution over today's 2-D digital TV broadcast infrastructure, the monoscopic color video has to be encoded using the standard MPEG-2 [21] tools currently required by the Digital Video Broadcast (DVB) project [2] in Europe, while the

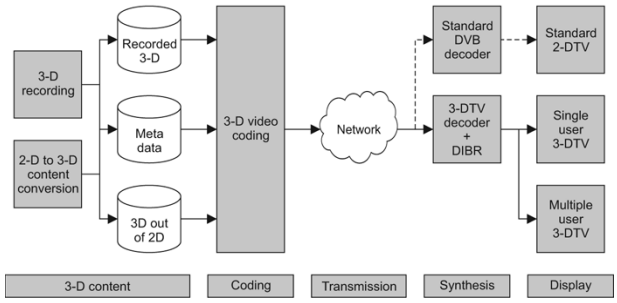


Fig. 2. 3-DTV signal processing and data transmission chain consisting of five functional building blocks: 1) 3-D content creation; 2) 3-D video coding; 3) transmission; 4) "virtual" view synthesis; 5) 3-D display.

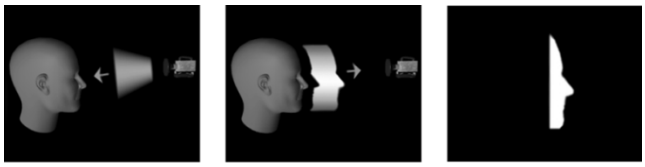


Fig. 3. Functionality of the Zcam active range camera. (a) An infrared light wall is emitted by the camera. (b) The reflected light wall carries an imprint of the captured 3-D scene. (c) The 3-D information is extracted by blocking the remaining incoming light with a very fast shutter (from [24]).

supplementary depth-images can be compressed using any of the newer, more efficient additions to the MPEG family of standards such as MPEG-4 Visual [22] or the latest Advanced Video Coding (H.264/AVC) [23].

To allow for an easier understanding of the fundamental ideas, the envisioned signal processing and data transmission chain of the outlined 3-DTV concept is illustrated in Fig. 2. It consists of five functional building blocks: 1) 3-D content creation; 2) 3-D video coding; 3) transmission; 4) "virtual" view synthesis; and 5) 3-D display.

A. 3-D Content Creation

A number of approaches are applicable for the creation of 3-D content. In one very appealing scenario, novel 3-D material is generated by simultaneously capturing video and associated per-pixel depth information with an active range camera such as the Zcam developed by 3DV Systems, Ltd. [24] or the NHK Axi-vision HDTV camera [25]. These devices integrate a high-speed pulsed infrared light source into a conventional broadcast TV camera, and they relate the time of flight of the emitted and reflected light walls to direct measurements of the depth characteristics of the 3-D scene (Fig. 3).

The main drawback of current *3-D cameras* is the fact that they are only fit for indoor use in studio environments and that they are not able to record more than relatively small-scale scenes (up to a few meters of depth). Thus, alternative approaches are required for the generation of 3-D data for larger scale, outdoor scenes. Here, the most promising concepts are based on the simultaneous capturing of multiview data using either traditional stereo cameras or synchronized multicamera systems (Fig. 4). Given several images of the spatial scenery, the 3-D geometry can be reconstructed by applying techniques from computer vision

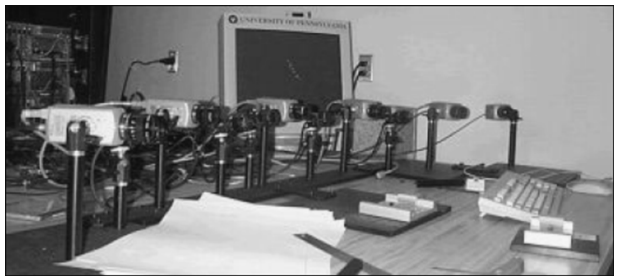


Fig. 4. The Penn State multicamera system. A cluster of up to six firewire cameras is used to generate depth information of a human participant in an immersive telepresence application (from [28]).

(CV) and photogrammetry [15], [26]–[28]. In general, most existing methods involve five basic steps: 1) geometric and photometric calibration of the individual cameras; 2) estimation of geometrical relations between the different views; 3) an optical flow or correlation-based search for corresponding points in two or more image planes; 4) localization of the corresponding 3-D space points; and 5) integration of the entire depth information into one or more camera reference frames.

Even with these novel "3-D capture" technologies at hand, it seems clear that the need for sufficient high-quality 3-D content can only partially be satisfied with new recordings. It will therefore be necessary—especially in the introductory phase of the new 3-DTV technology—to also convert already existing 2-D video material into 3-D using so-called "structure from motion" algorithms. On principle, such (offline or online) methods process one or more monoscopic color video sequences to: 1) establish a dense set of image point correspondences from which information about the recording camera as well as the 3-D structure of the scene can be derived [26], [27], [29], [30] or 2) infer approximate depth information from the relative movements of automatically tracked image segments [31].

B. "Virtual" View Synthesis

DIBR is defined as the process of synthesizing "virtual" views of a real-world scene from still or moving images and associated per-pixel depth information [32], [33]. Conceptually, this novel view generation method can be understood as a two-step procedure: At first, the original image points are reprojected into the 3-D world, utilizing the respective depth values. Thereafter, these intermediate space points are projected into the image plane of a "virtual" camera located at the required viewing position. The concatenation of reprojection (2-D to 3-D) and subsequent projection (3-D to 2-D) is usually referred to as "3-D image warping" in the computer graphics (CG) literature.

1) The "Virtual" Stereo Camera: Building on the described 3-D image warping concept, the synthesis of stereoscopic images can be realized through the definition of two "virtual" cameras—one for the left-eye and one for the right-eye. With respect to the original (reference) view, these two cameras are symmetrically displaced by half the *interaxial distance* t_c (Fig. 5). To establish the zero parallax setting (ZPS), i.e., to choose the *convergence distance* Z_c in the 3-D

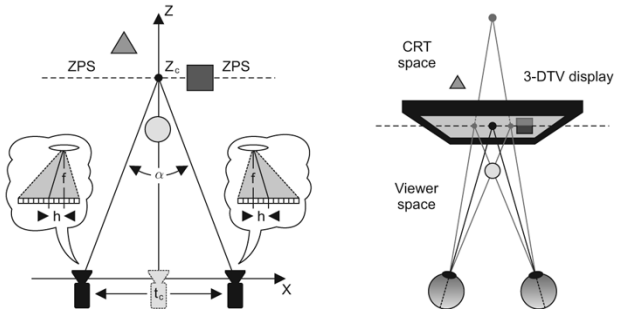


Fig. 5. A "virtual" stereo camera with shift-sensor convergence and the respective 3-D reproduction on an (auto)stereoscopic 3-DTV display. Scene parts that lie further away than the convergence distance Z_c are visualized behind the screen in CRT space, areas closer than Z_c are reproduced in front of the display in viewer space.

scene, the CCD sensors of the parallel positioned cameras are translated by a small shift h relative to the position of the lenses. In real stereo cameras, this shift-sensor concept is usually prefered over the alternative "toed-in" approach, because it does not introduce keystone distortions and depth-plane curvature in the stereoscopic imagery [34], [35]. When implemented with DIBR, it has the additional advantage that all the required signal processing (e.g., the antialiased resampling of the "virtual" left- and right-eye views) is purely one-dimensional [14].

The 3-D transformations that transfer the original image points (u, v) to their new locations (u_l, v) (respectively, (u_r, v)) in the "virtual" left- and right-eye views can easily be derived from the perspective projection model. They result to

$$\left.\begin{matrix} u_l \\ u_r \end{matrix}\right\} = u \pm \frac{\alpha_u t_c}{2}\left(\frac{1}{Z} - \frac{1}{Z_c}\right) \tag{2}$$

where α_u is the focal length f of the reference (center) camera expressed in multiples of the pixel width w [36]. (A detailed derivation of these equations can be found in [14].)

2) Flexibility of 3-D Reproduction: Table 1 shows how the 3-D reproduction that results from (2) is influenced by the choice of the three main system variables, i.e., by the choice of the interaxial distance t_c, the focal length f of the original camera, and the convergence distance Z_c. The respective changes in parallax, perceived depth, and object size are qualitatively equal to what happens in a real stereo camera when these system parameters are manually adjusted. For example, an amplification of the interaxial distance t_c leads to an increase in parallax. This, in turn, leads to an increase in perceived depth without, however, affecting the perceived object size. A magnification of the focal length f, on the other hand, not only leads to an increase in parallax and perceived depth, but also increases the perceived object size. Finally, changing the convergence to a larger distance Z_c decreases the parallax without affecting the overall perceived depth and object size. However, the 3-D scene will appear to be shifted forward on any autostereoscopic 3-DTV display.

The main advantage of a "virtual" stereo camera over a real one is the fact that the stereoscopic system parameters must not be defined at *capture time* (except for the focal length f). Rather, they can be optimized at *display time* to customize the resulting 3-D reproduction for a specific viewing condition. This allows the users to adjust the depth percept according to individual preferences [14].

Table 1
Influence of "Virtual" Stereo Camera Parameters (After [37])

Parameter	+/–	Parallax	Perc. depth	Obj. size
t_c	+	Increase	Increase	Constant
	–	Decrease	Decrease	Constant
f	+	Increase	Increase	Increase
	–	Decrease	Decrease	Decrease
Z_c	+	Decrease	Shift (forward)	Constant
	–	Increase	Shift (backwards)	Constant

The effect of interactively changing the parallax

$$P(Z) = u_r - u_l = \alpha_u t_c \left(\frac{1}{Z_c} - \frac{1}{Z}\right) \tag{3}$$

by means of a variation of the interaxial distance t_c is visualized exemplarily in Fig. 6. The magnified clippings from two "virtual" stereoscopic images show that for $t_c = 24$ mm only a rather low amount of parallax—and thus a relatively small depth effect—results for this specific scene. For $t_c = 48$ mm the parallax values are exactly twice as large as in the first case and, correspondingly, the perceived depth impression is also enlarged approximately by this factor.

3) The Disocclusion Problem: An inherent problem of the stereoscopic view synthesis concept is due to the fact that areas, which are occluded in the original view, might become visible in any of the "virtual" left- and right-eye views, an event referred to as *exposure* or *disocclusion* in the computer graphics (CG) literature [32], [33]. The resulting question is how these disocclusions should be concealed during the "virtual" view generation, as information about the previously occluded areas is neither available in the monoscopic color video nor in the accompanying depth-image sequence.

In the absence of any additional image data, there are only two basic options. One can try to: 1) replace the missing image areas (holes) during the view synthesis with "useful" color information or 2) preprocess the depth information in a way that no disocclusions appear in the "virtual" views. One option, for example, is to treat exposures as regular magnifications and to close them with a simple foreground/background interpolation during image resampling [Fig. 7(a)]. The weak point of this approach is that it can result in annoying "rubber-sheet" artifacts [33], i.e., visible color gradients appearing between objects located at different levels of depth. Visually less perceptible impairments can be achieved by preprocessing the per-pixel depth information with a 2-D Gaussian low-pass filter such that large discontinuities are effectively smoothed out [Fig. 7(b)]. While this obviously leads to some geometric distortions in the stereoscopically reproduced 3-D scene, the perceived visual quality was found to be acceptable for test images with a rather low amount of parallax [38].

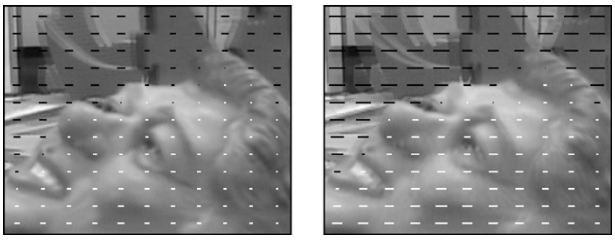


Fig. 6. Changing the interaxial distance t_c of a "virtual" stereo camera. Magnified clippings from two "virtual" stereoscopic images synthesized with less (respectively, more) parallax, i.e., for a smaller (respectively, a larger) perceived depth effect. (The black vectors indicate positive parallax values, the white vectors visualize negative parallax values.)

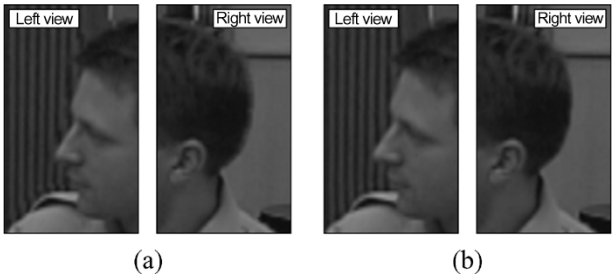


Fig. 7. "Virtual" view synthesis examples. Two different approaches to conceal the disocclusions between foreground (the person's head) and background (the wall) are compared. (a) Simple foreground/background color interpolation. (b) Prefiltering (smoothing) of the per-pixel depth information using a 2-D Gaussian low-pass filter.

For larger parallax values or for the more excessive disocclusions that would result from supporting head-motion parallax (HMP) viewing on single or multiview 3-DTV displays, higher quality synthesis results can only be achieved when additional information about the occluded areas is generated already during the creation of the 3-D content. One possible solution is provided by the so-called layered depth-images (LDIs). As an enhancement to the basic depth-image, LDIs allow to store more than one pair of associated color and depth values for each pixel of the original image, with the number of layers typically depending on the scene complexity as well as the required synthesis quality [39]. With such a 3-D data representation format, additional hidden layer information can be used to fill-in the exposed areas in the "virtual" left- and right-eye views. Because the data in the extra layers is usually quite sparse, the overhead required for storage and transmission in general is not that big. Another step toward photorealistic synthesis quality was recently proposed by Zitnick *et al.* [15]. To deal with *mixed* boundary regions, i.e., object border pixels that received contributions from both foreground and background colors, they transmit an additional opacity map (alpha channel) that is extracted using a Bayesian matting technique [40].

C. 3-D Video Coding

We believe that future 3-DTV services can only become commercially viable (at least in a conventional broadcast TV environment) when they do not require excessively more bandwidth and storage capacities than regular 2-D digital TV programming. Thus, an important requirement for any 3-D data representation format is that it can be encoded in a highly economical manner.

In order to analyze the suitability of the different MPEG standard technologies for efficient compression of typical depth-images, a comparative coding experiment was performed [14]. The test group consisted of three video codecs: 1) the MPEG-2 reference model codec (TM-5); 2) a rate-distortion (R/D) optimized MPEG-4 Visual codec developed at Fraunhofer HHI; and 3) the R/D optimized H.264/AVC reference model codec (v6.1a). The compression results for the two test sequences "Interview" and "Orbi" are shown in Fig. 8 for typical broadcast encoder settings, i.e., for a GOP (group of pictures) length equal to 12 with a GOP structure of IBBPBBP. . ., by means of rate-distortion curves over a wide range of different bitrates.

The provided results show that H.264/AVC as well as MPEG-4 Visual are very well suited for the compression of per-pixel depth information (with H.264/AVC being even more efficient). If a bitrate of 3 Mbit/s is assumed for the transmission of MPEG-2 encoded monoscopic color information over a conventional TV broadcast network (e.g., DVB-T), it can be seen from the two graphs that the accompanying depth-images can be compressed to target rates significantly below 20% of this value. This makes 3-DTV possible with only a minor overhead compared to today's 2-D digital TV. To give an example, H.264/AVC compression of the "Interview" sequence at 105 kbit/s (3.5% of 3 Mbit/s) still leads to a very high PSNR of 46.29 dB. For the slightly more complex "Orbi" scene, this value can still be reached at a bitrate of approximately 184 kbit/s (6.2% overhead).

D. Transmission

To transmit the 3-D data in a backward-compatible manner over the conventional 2-D digital TV broadcast infrastructure, all video sequences (i.e., the monoscopic color video, the associated per-pixel depth information, as well as any other possible enhancement layers; cf. Section II-B) have to be synchronized to each other and multiplexed together into an MPEG-2 transport stream (TS) [41], which can then be distributed to the users via cable (DVB-C), satellite (DVB-S), or terrestrial (DVB-T) transmitters [2]. For 3-DTV transmissions over the Internet, synchronization of multiple media streams can be realized efficiently by relying on state-of-the-art delivery control mechanisms such as the Real-Time Transport Protocol (RTP) [42].

III. Autostereoscopic 3-DTV for Single and Multiple Viewers

When finally entering the mass market, 3-DTV will have to compete with viewing-comfort and visual-quality standards set by advanced television systems providing superb high-definition images on wide-screen full-color and high-contrast display devices. Hence, the question arises whether there will be an extra benefit of 3-DTV compared to 2-DTV in terms of a more intense and satisfying television experience or of special possibilities to interact with the programming content, so that consumers will like it, use it, and be willing to pay for it.

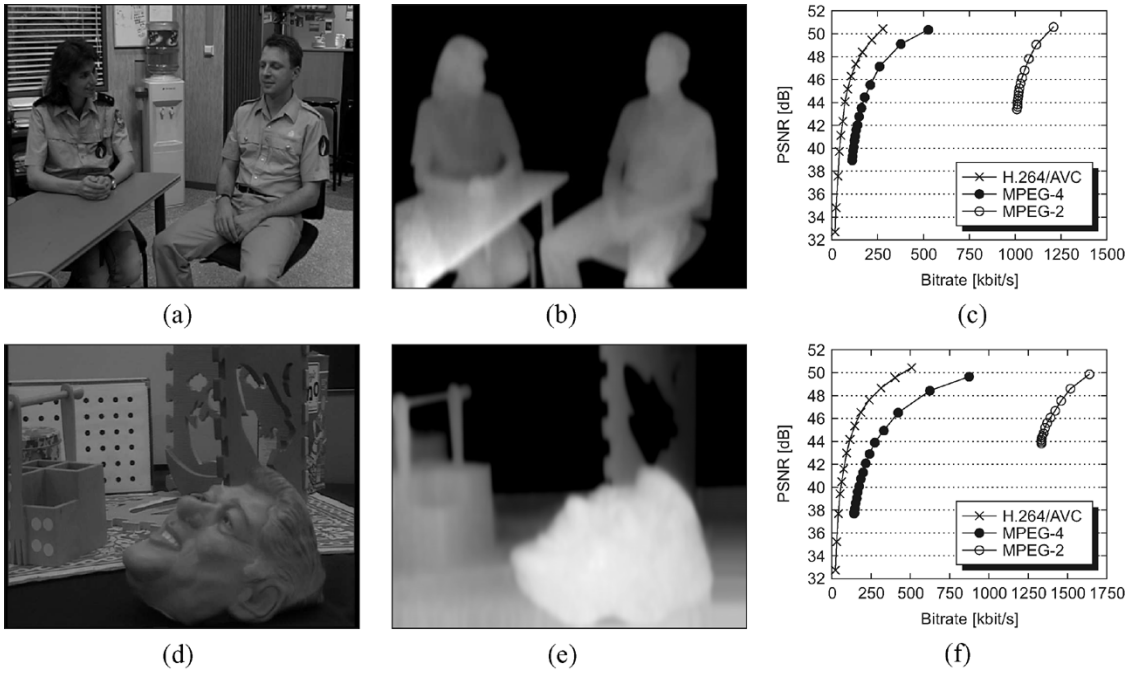


Fig. 8. Results of compressing the depth-images of the "Interview" and "Orbi" test sequences with three different MPEG codecs. (a), (d) Monoscopic color video. (b), (e) Per-pixel depth information. (c), (f) Coding results shown as rate-distortion curves.

There is no doubt about 3-D superiority in numerous professional applications ranging from robotics to medicine. But when it comes to news broadcasting or televised entertainment, it is by far not obvious why binocular depth information should add anything essential to the message or enhance the enjoyment of the show. This question was addressed in various human factors studies carried out in the 1990s, where simulated advanced 2-DTV standards served as a reference (see e.g., [20], [43], [44]). From a synopsis across collected user responses, a clear preference for 3-D emerged: in direct comparison with 2-DTV, the immediate impact of 3-DTV was felt more intense and satisfying, identical scenarios were rated more appealing and interesting, and under forced choice conditions there was an exceptionally clear-cut decision in favor of 3-D. Moreover, about the same viewing distance with regard to the picture size was preferred for 2-D and 3-D, suggesting a good possibility of coexistence between high-definition 2-DTV and 3-DTV in terms of the basic imaging parameters.

The bad news, however, is that these encouraging results were obtained under "otherwise identical" viewing conditions, meaning that there were no special restrictions for 3-D versus 2-D viewing. This outcome has significant consequences for 3-D display developers: future 3-DTV consumers want 3-D appliances differing from advanced 2-D sets in nothing but the added feature of rendering binocular depth cues—i.e. they want a 3-D display that can be viewed without special stereo glasses, with almost no limitations on the seating position, suitable for group viewing, and with a picture quality (resolution, size, contrast, etc.) comparable to advanced 2-DTV.

Various attempts ranging from direct-view 3-D monitors to large-screen 3-D projection displays were made, with the result that, up to now, no 3-D display technology exists that can fulfill all user requirements at the same time. Some technologies provide high-resolution stereo images at relatively unrestricted viewing positions—however, for a single viewer only. Displays for group viewing create limited viewing zones, generally coupled with annoying visual effects (image flipping) when the user moves. Moreover, with direct-view displays for group viewing, image resolution is reduced to a level of $1/N$ of the base 2-D display technology used (for each view of an N-view system) and contrast is downgraded by interview crosstalk. Crosstalk also affects the depth perception; too much crosstalk will prevent the user from beeing able to see the 3-D effect (the perception threshold for crosstalk is at 0.3% [44]). In the following, we will briefly overview the latest trends in glasses-free (autostereoscopic) 3-DTV display development that appear suitable to meet the essential user demands. Interested readers find detailed treatises on the foundations of 3-D displays in the works of Valyus [45] and Okoshi [46]. A comprehensive survey of display solutions is given by [47].

A. *Personal TV*

In the near future, our experience of television will drastically transform. What is currently a broadcast linear entertainment system mainly used for relaxation is becoming an on-demand, participatory, and interactive communications platform. Technologies originally developed for the PC, such as random-access disks and high-speed connections to the Web are increasingly used in connection with television appliances, while advanced video coding techniques such as H.264/AVC and WMVHD allow streaming of high-resolution media content over the Internet. The

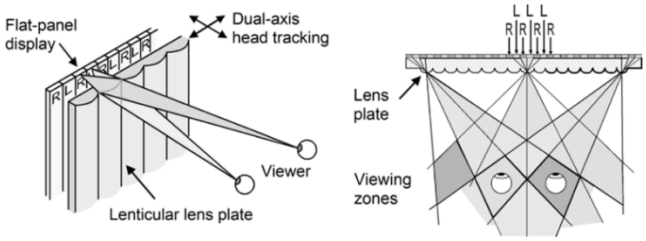


Fig. 9. Principle of operation of a direct-view lenticular lens display with head tracking. L and R denote corresponding columns of the left- and right-eye images. The lens pitch is slightly less than the horizontal pitch of the pixels (depends on the design viewing distance).

emerging convergence of computing and television is actively supported by renowned companies like Microsoft Corporation with their Windows Media Center as well as companies providing technologies and video programming for interactive and individualized television [3], [4]. Hence, it is expected that in the not-so-distant future there could be a significant market for personal (2-D and 3-D) TV appliances, used interactively by single viewers the same way as the PC is used by a single person.

Stereoscopic 3-D displays provide different perspective views to the left and right eye. Like in natural vision, the differences in perspective are immediately used by the visual system to create a vivid, compelling, and efficient sensation of depth in natural and computer generated scenes. There are two basic concepts in the underlying 3-D display technologies. One relies on special user-worn devices, such as stereo glasses or head-mounted miniature displays, in order to optically channel the left- and right-eye views to the appropriate eye. Another concept integrates the optical elements needed for selective addressing of the two eyes in the remote display device, hence allowing free 3-D viewing with the naked eyes. In general, such autostereoscopic 3-D displays are more comfortable to the viewer, and they are a must for future 3-DTV.

Various practical concepts have been proposed to create high-resolution stereoscopic video images which can be watched by a single viewer without special stereo glasses from a range of viewpoints. One basic approach is to multiplex the two perspective views, e.g., on a pixel-by-pixel basis, and to use an array of microoptical elements in order to separate the views again and to make them visible to the left and right eye, respectively. Another approach employs separate display panels for the two views; in this case, macro-optical elements including large mirrors and field lenses are used to direct the light emitted by the two panels exclusively to the appropriate eye. Both approaches are inherently interactive, since they make use of head tracking in order to sense the viewer's current eye position and to adapt the optical paths accordingly. Head tracking allows individual users to move freely in front of the display without losing "contact" to the 3-D display content.

1) Multiplexed Personal 3-DTV Displays: Fig. 9 illustrates the working principle of the pixel-multiplex approach in connection with an array of cylindrical lenslets used to separate the two views. The two stereo views are presented simultaneously, with two columns of pixels (one for the left- and one for the right-eye image) behind a lenslet. There is one diamond-shaped main viewing zone with optimal stereo separation and a number of adjacent side lobe zones. Side lobes result from those rays of light travelling not directly through the front most lenses but through neighboring lenses. In the side lobes, depth may be inverted if the left and right images are not seen with the proper eye. When traversing from a viewing zone to the adjacent one, the display appears more or less darkened (depending on the viewing distance), since the spaces between the pixel cells covered by the black matrix create black zones between the individual views. In order to avoid such interferences, the observer's head position should be constantly sensed and the lenticular sheet should be shifted in relation to the pixels, mechanically or digitally, to track head movements.

Fig. 10. Prototype high-resolution personal 3-DTV monitor showing 3-D video content and a 2-D or 3-D enhancement window (content reproduced by permission of unterwasser-photographie.de, Berlin).

Because of the horizontal selectivity of the lens array, the color-filter stripes of the display panel must be aligned one above the other. Otherwise, the viewer would see separated color components from different viewpoints. Unfortunately, in available flat-panel displays the RGB subpixels are arranged horizontally in a line. The "simple" solution is to operate the display in portrait mode. The landscape mode requires a technically more demanding subpixel multiplexing scheme and a significant reduction of the lens pitch (cf. Fig. 11). Moreover, it must be taken into account that the usable horizontal aperture of the subpixel cells of an RGB triplet is less than one third of the aperture in the vertical direction. Overall, these unfavorable conditions put very high demands on the production and alignment tolerances of a lenticular lens display operating in landscape mode.

A fully operational prototype based on a high-resolution 21.3-in LCD panel with $1200 \times 1600 \times$ RGB pixels was built at the Fraunhofer HHI. As shown in Fig. 10, the display is operated in portrait mode, thus providing almost perfect stereo separation with less than 2% crosstalk. Due to the dual axis head tracking controlled by a 120-Hz eye-position tracker,

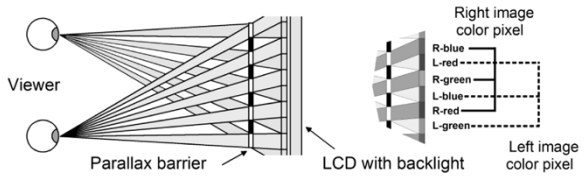


Fig. 11. Principle of a parallax-barrier 3-D display and grid versus pixel arrangement for standard landscape-format direct-view panels with vertical color-filter stripes.

there is an unrestricted viewing zone for a single user ranging from 400 to 1100 mm in the frontal direction and 500 mm in the vertical direction. Left/right head motion is tracked in a range of 60°.

Since normal TV programming applies the landscape image format, the display gives room for additional information and interactive elements that do not interfere with the video screen (as opposed to the overlay techniques used in current interactive TV). Such a split screen concept offers various design options for personal 3-DTV applications. For example, the extra space could be used to display 2-D content, such as the electronic program guide (EPG) and related Web frames enhancing the show. In this case, the lenticular lens array would be proportioned to the 3-D video area.

A different approach to high-resolution flat-panel 3-D displays developed at Dresden University, Dresden, Germany, employs an array of microprisms in place of the lenslets, in order to deflect the rays of light passing through the alternate pixel columns (cf. Fig. 9) to the left and right eye, respectively. A commercially available product is offered from SeeReal Technologies [48]. Other well-elaborated approaches use a so-called parallax barrier array placed at a distance in front of the imaging panel. Due to parallax effects, one eye can see only pixel-columns of one view through the barrier's slit openings, while pixels belonging to the other view are occluded by opaque stripes (Fig. 11).

A general disadvantage of 3-D displays using microoptical arrays according to the basic principle illustrated in Figs. 9 and 10 is the fact that only one-half of the spatial resolution of the native panel is available for each of the two views. This problem can be overcome by time-sequential display of the two views in connection with a switchable parallax barrier. For example, in the first phase only the odd-numbered pixel columns of the two views are sent to the imaging display (where they are presented on alternate columns), followed by the even-numbered columns in the second phase. The left and right-view pixel columns are shifted by one pixel position when switching between the phases, and the parallax barriers are inverted accordingly, in order to direct the pixels to the intended eye. An approach developed at New York University, New York, by Perlin *et al.* suggests applying three phases [6]. This allows separating the visible image stripes by black spaces, which allows for some registration errors in the system. Hence, this concept reduces the required precision of head tracking.

2) Field-Lens Based Personal 3-DTV Displays: In various optical instruments, a field lens is placed at the locus of a real (aerial) image in order to collimate the rays of light passing through that image, without affecting its geometrical properties. Various 3-D display concepts use a field lens to project the exit pupils of the left and right-image illumination systems into the proper eye of the observer. The result is that the right-view image appears dark to the left eye, and *vice versa*. This approach avoids all the difficulties resulting from small registration tolerances of pixel-sized optical elements and it provides full resolution images to the viewer's left and right eye.

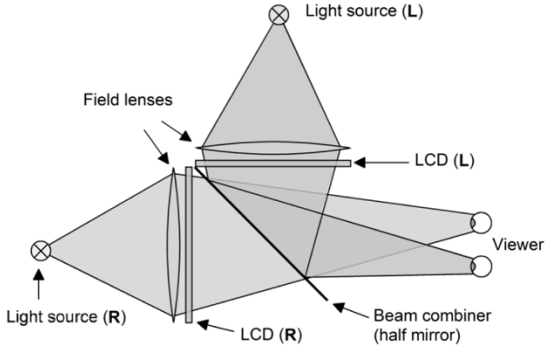


Fig. 12. Schematic view of a dual-LCD field lens display.

The principle of operation is shown in Fig. 12 [49]. The left and right views displayed on distinct *LC* panels are superimposed by a beam combiner. Field lenses placed close to the LCDs direct the illumination beams into the proper eye. For head tracking, the position of the light sources is steerable.

A different concept using a high-resolution stereoscopic video projector has recently been developed at Fraunhofer HHI [50]. This approach allows variable positioning of the (virtual, aerial) image plane, where the stereo pair is perceived by the viewer. Hence, it is possible to make interactive elements floating in space within the user's arm reach, so that they can be "pressed" with the finger like a natural button. Moreover, it is possible to steer the distance of the image plane according to the distance of a currently observed object in the displayed 3-D scene. This "trick" allows the user to accommodate his/her eyes on variable object distances, like in natural vision, making the display particularly comfortable to view.

Fig. 13 illustrates the basic optical concept, and Fig. 14 shows a display prototype providing a stereo resolution of $2\times1600\times1200\times$RGB pixels. Head tracking is achieved by a mechanically moving stage, making the optical path follow the viewer's eye position. Due to the optical principle, the perceived stereo image is very bright (more than 800 cd/m^2 with our prototype) and the optical system perfectly separates the left and right views (zero image crosstalk). The Scottish company IRIS-3D has recently developed a display based on a similar principle of operation, where large concave mirrors are used in place of the collimating Fresnel lens in the HHI display [51].

B. Multiple-Viewer 3-DTV

Three-dimensional TV displays for more than one simultaneous user watching TV require that left- and right-eye views are channelled to multiple eye pairs. This can be achieved

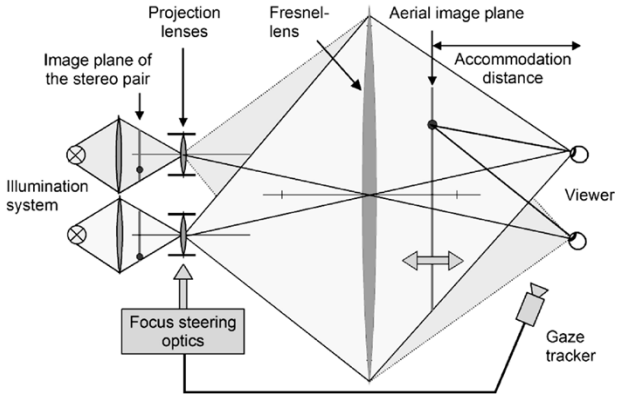


Fig. 13. Top view of the principle of operation of a field-lens based personal 3-DTV display with variable accommodation distance. The viewer accommodates on the aerial image plane and perceives a floating 3-D image. The Fresnel lens images the exit pupils of the stereo projector at the locations of the left and right eye, respectively, and hence separates the constituent stereo images. The aerial image plane is dynamically aligned with the current position of the viewer's fixation point.

Fig. 14. Prototype autostereoscopic field-lens based Personal 3-DTV display. The display creates high-resolution aerial 3-D images floating 25 cm in front of the screen (a special 30-in filter panel) within the user's arm reach. A video-based finger tracker is embedded in the desktop at a location corresponding with the aerial image.

either actively by multiuser head tracking or, passively, by displaying multiple adjacent views covering the range of intended viewing positions. If microoptical arrays are used for stereo separation, it is possible to use the repeated viewing zones created by side-lobes, in order to accommodate multiple viewers.

1) Multiple-Viewer 3-DTV With Head Tracking: At present, there are only a few multiple-viewer 3-DTV concepts based on multiuser head tracking and even fewer have been implemented so far. In principle, the dual-view LCD field lens display illustrated in Fig. 12 can be used by multiple viewers, if multiple pairs of left/right illuminators are used. While in this concept two large field lenses serve to image the exit pupils of the display's backlighting system at the viewers' eye locations, a novel approach under development at the De Montfort University, Leicester, U.K., substitutes the function of the large lenses and light sources by an array of multiple small optical elements and illumination sources [52]. The steering optics allow to illuminate the entire screen area and to direct the emitted light beams

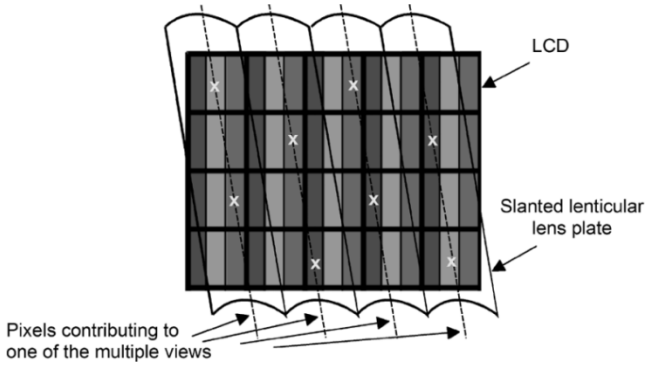


Fig. 15. Principle of the slanted-raster multiview approach (seven-view display, after van Berkel and Clarke [53]). From a particular viewing angle the eye will see points of the display panel sampled along the dashed lines. Color subpixels marked by an "x" belong to the same perspective. When the eyes move from one perspective to the adjacent one, the overall brightness of the display is approximately maintained and a gradual fading between the views is observed.

to the viewers' eye positions. Several illumination sources can be lit simultaneously, producing multiple exit pupils for the multiple viewers.

2) Multiplexed 3-DTV Displays for Multiple Viewers: Multiview 3-DTV displays present more than two perspective views across the viewing field. Hence, according to Fig. 9, multiple pixel columns, each belonging to one of the multiple views, must be accommodated behind the microoptical elements (lenslets or slit openings of a parallax barrier). An obvious shortcoming of this approach is that just a fraction of the native horizontal resolution of the basic display panel is available for the stereo images, while the vertical resolution remains unchanged.

This problem is neatly solved by setting the lenslet or slit array at an angle in relation to the display panel (Fig. 15). Depending on the number of views and the multiplex scheme applied, this concept gives developers a way to control the resolution tradeoffs in the horizontal and vertical direction. Moreover, when the eyes move to adjacent perspectives there is no such abrupt change (flipping) as observed in the traditional design, where the lenses are vertically aligned with the columns of the display panel. On the other hand, with a slanted microoptical raster it is not possible to perfectly separate adjacent perspectives. The partial visibility of pixels belonging to adjacent views produces crosstalk limiting the usable depth range and contrast of the display.

Multiplexed multiview displays inherently require very high resolution imaging panels, since only a fraction of the original pixels is available for each of the distinct views. For example, a high-definition ten-view 3-DTV display would require ten times the number of pixels of an ordinary HDTV display panel, what is far beyond current technologies. TV content production requires either an array of cameras, simultaneously capturing multiple views of a scene, or dedicated real-time "virtual" view synthesis (cf. Section II-B) or similar view interpolation techniques [54].

3) Multiview 3-DTV Projection: Multiple video projectors can be used to increase the native resolution of the 3-DTV display. The idea dates back to the early 1930s, when researchers at Bell Telephone Labs proposed to project

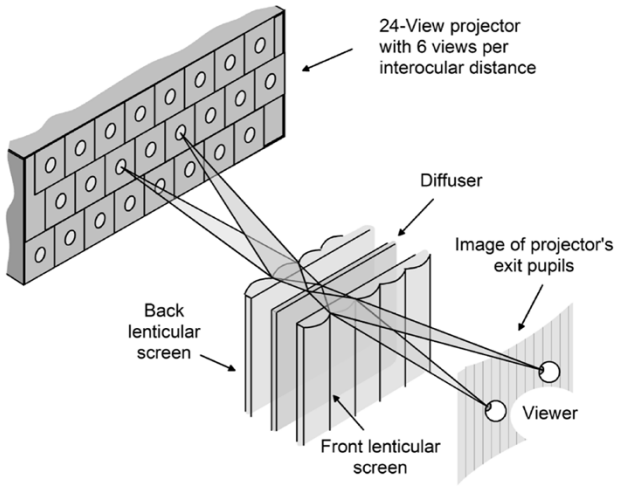


Fig. 16. Principle of a 24-view rear projection display with six views per interocular distance. The projection side lenticular screen focuses the images in the form of vertical stripes onto a translucent diffuser. The lenses of the viewer-side screen map the vertical image stripes to specific loci in the viewing zone—quasi-mirroring the initial path of light during rear projection. At the design viewing distance, images of the multiple exit pupils of the projector form an array of seamlessly adjacent viewing zones for the different views. Since the light beams scattered by the diffuser also pass through a set of adjacent lenslets, there are several adjacent view fields for multiple viewers (not shown in the diagram).

multiple perspective views onto a retro-reflective lenticular lens screen (a screen that reflects the incoming beams of light back to their origin [46]). Böner of Fraunhofer HHI optimized the projector arrangement and screen design (lenses, screen curvature, and diffuser material) for side-lobe viewing by multiple viewers and proposed a special microstructuring of the front surfaces of the lenslets in order to reduce annoying reflections of the projectors exit pupils (showing up as extended light bands running horizontally across the screen). A prototype 100-in screen with about 3000 lenslets was fabricated and used in connection with a special 24-view slide projector [55], [56]. For rear projection with a sandwiched double lenticular screen (Fig. 16) a special diffuser layer with embedded acrylic microbullets was manufactured.

Recently, researchers at Mitsubishi Electric Research Labs, Cambridge, MA, presented the first real-time end-to-end 3-DTV system based on the multiview projection approach [8]. This prototype was implemented with rear-projection and front-projection lenticular screens and simultaneously displayed 16 stereoscopic video images viewable from multiple viewpoints.

In general, multiview 3-DTV projection systems require a very large number of views with only slight differences in perspective. Otherwise, the change in perspective when moving creates noticeable "jumps" (image flipping) [57]. Crosstalk between adjacent views can soften this effect. User tests with an experimental 100-in 3-D-HDTV rear-projection display showed that 6–18 views per interocular separation (65 mm) are required to make flipping "imperceptible" or at least "not annoying" (depending on interview crosstalk) [44]. Hence, if for example the width of the individual viewing zones is 650 mm at the design viewing distance, a dense array of 60–180 video projectors is required for an adequate multiview projection 3-DTV display. Further issues to be solved include the development of pupil-forming optics with sufficiently large (preferably square) apertures as well as methods that suppress possible Moiré distortions due to interferences between the regular pixel patterns of the light valves (projection LCDs and DMDs) and the regular array of the lenslets on the screen. Special diffusers are required that provide a best visual balance between homogeneous brightness across the screen and interview crosstalk.

IV. Viewer Interaction

Interactive television is, essentially, video programming which incorporates some kind of interactivity for enhancement of the programming. The interactive elements in use include icons, banners, labels, menus, interface structures, open text fields in which users can insert their e-mail addresses, forms to fill out in order to buy a product, or commands to retrieve and manage video streams and graphics on a related Web page. Usually, these elements will partly cover the broadcast. Moreover, viewers will be able to control camera angles during live events, select which commercials they want to watch (and interact with), and control a selection of choices content producers provide as part of the broadcast [58].

For navigation, viewers typically use the buttons on their remote control, type on a wireless keyboard, or use a trackball interface integrated in the keyboard. In a more futuristic setting allowing interactive scene walkthroughs, a wireless gamepad lets the user navigate through the captured environment [59].

Recently, various machine-vision based techniques have been developed capable of sensing people and interpreting their gestures. These techniques were employed in order to enhance traditional human–computer interaction, particularly in connection with 3-D graphical user interfaces and autostereoscopic 3-D displays [60]. Hence, there is a good chance that future users of interactive 3-DTV will benefit from input devices allowing them to use their natural interaction modalities such as speech and gesturing.

Hand gestures could be useful for selecting interactive objects on the TV screen by pointing to them (deictic indications) or touching them with a finger, to move irrelevant objects aside, or to indicate the direction of simulated ego-motion in interactive scene walkthroughs (viewpoint and trajectory, viewing angle). Facial gestures convey emotion, belief as well as intentions and may be interpreted in order to sense the attentiveness of a viewer to the current TV content. Gaze direction immediately indicates the user's focus of attention and the level of attraction of specific screen content; wandering gaze or gaze to unrelated items signal a loss of engagement, and may signal a desire to end the interaction [61]. Head nods and shakes could signal affirmative reactions to a broadcast content. Moreover, head movements are a natural means to inspect a single object from different perspectives or to find out the spatial relations of multiple objects in a 3-D scene.

Apart from using gestures as input signals controlling the device, one can imagine future game shows *à la Fahrenheit 451*, where gaze, head movement, arm movement, and body stance are part of the interaction of remote participants. Moreover, video analysis of the eye and face region could be used to identify the viewer with subsequent profiling of the broadcast content and interface structure according to individual preferences. In the following, we will briefly overview some recent solutions for sensing interactive gestures, focusing on developments of our labs.

A. Head Tracking

There are two matured vision-based head tracking systems, one developed by IBM Almaden for human–computer interaction (the BlueEyes tracker) and the other one developed by Seeing Machines, an Australian company specialized in real-time head and eye-gaze tracking, for driving studies in test vehicles (the faceLAB system) [62]. The BlueEyes tracker is a dual-light source head tracker using the dark versus bright pupil effect [63]. One set of infrared LEDs mounted close to the camera axis creates a bright pupil image (red eye effect). A second off axis IR source produces a dark pupil image. The two light sources are switched on and off alternatively for subsequent video frames. Hence, the pupils are marked as bright spots in the difference image. The faceLAB system is a robust and flexible stereovision solution, based on a head model adapted to the facial features of a user. After individual calibration to the user, the system extracts the location of facial features such as the eyebrows, pupils, iris, eye corners, mouth, and nostrils, and estimates the pupil positions, head pose, and gaze direction.

At Fraunhofer HHI, we have developed a prototype high-speed, high-precision stereo video head tracker. Based on the edge orientation matching method proposed by Fröba and Küblbeck [64], the tracker initially extracts the edges in the video image and analyzes the orientation patterns using a face detection model. The detection model was obtained from a statistical analysis of edge orientations in a large set of facial and nonfacial training sample images. Basically, it determines whether a certain edge angle at a certain pixel position is more likely to occur in a face or nonface image. For example, in the eyes and mouth regions, edges run prominently in the horizontal direction while the nose area is characterized by vertical edges. The search is performed in a course-to-fine procedure on three levels of a resolution pyramid. Possible face candidate positions are classified by a neural network with two output targets (face and nonface). After the eye regions have been found, the current user's individual eye patterns are used for tracking with a multiple-pattern, adaptive block matching algorithm (implemented in assembler using the Pentium IV SSE2 multimedia instruction set). The initial reference eye patterns are scaled using an affine transformation corresponding to six different user-to-camera distances. The resulting 12 reference eye patterns are finally used by the head tracker to find the eyes in the current live video images.

The two-stage approach was applied to speed up and simplify tracking. With high-speed stereo cameras the tracking rate is 120 Hz at a precision of $3 \times 3 \times 10$ mm (x, y, and z positions of both pupils). The baseline of the stereo setup may be as large as 800 mm, so that the cameras can be mounted to the left and right side of a 30-in 3-D display. Fig. 17 shows that the algorithm keeps track of an individual user's eyes, even when other faces appear in the camera image.

Fig. 17. Tracked eye positions in the presence of other faces in the camera image. The size and format of the pixel array keeping the individual eye pattern can be adapted to get optimal tracking results.

Current challenges associated with vision-based head tracking can be attributed to the following factors: Variation of the captured head images due to the current head orientation and the occurrence of (self)occlusions (pose problem); presence or absence of feature components like glasses and facial hair; variation of facial expressions; fluctuation of environmental conditions, such as changes in the illumination intensity and the position of illumination sources; detection and precise tracking of multiple faces in images (required for a head-tracked multiple-viewer 3-D display).

B. Gaze Tracking

The gaze tracking algorithm developed at Fraunhofer HHI applies a special cornea-reflex method based on the approach described by Young and Sheena [65]. It senses the user's current point of fixation (lines of sight of both eyes) at a rate of 50 Hz with a single stationary camera (Fig. 18). Due to the wide-angle optics in combination with a 1280 $\times$ 1024 pel resolution camera, the user may move in a range of 300 $\times$ 300 $\times$ 300 mm. A two-step calibration procedure requires minimal individual calibration for the end-user. An alternative solution based on the evaluation of various facial features, such as the position of the iris and pupil with regard to the corners of the eye, is offered by Seeing Machines [62].

We have opted for the cornea-reflex method because of its high precision and stability and because it is nonintrusive. The viewer's eyes are illuminated with low-intensity infrared light. An array of LEDs is mounted at a distance from the optical axis of the gaze camera. As a result, the pupils appear as black elliptical regions. The centers of the pupils and the reflections of the light from the cornea are found rapidly due to a new algorithm. There is a monotonic geometrical relationship between the vector pointing from the center of the pupil to the center of a light reflection (eye vector) and the

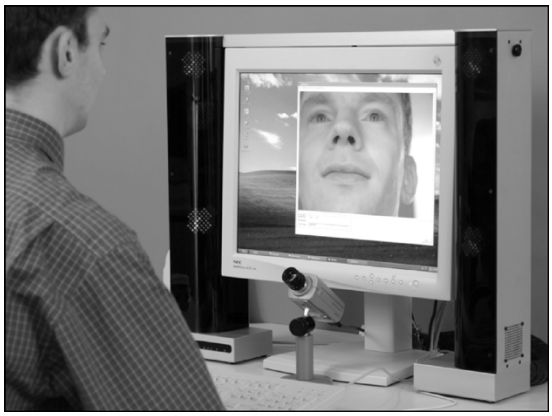


Fig. 18. The gaze tracker camera captures a high-resolution image of the user's eye. Switchable infrared illuminating LEDs to the left and right of the monitor create reflections on the eye surface; their position with respect to the pupil position is evaluated for sensing the viewer's gaze direction.

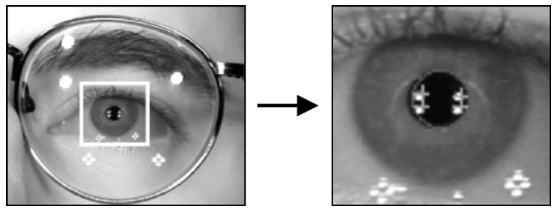


Fig. 19. The eye image captured by the gaze tracker (left) and an enlarged part of the eye showing the pupil and light reflections on the cornea (right). Note the distracting reflections on the user's corrective glasses.

Fig. 20. The Fraunhofer HHI stereo hand tracker module and the graphical user interface showing the segmented hand, the fingertips found, and the tracks of palm and index-finger movement.

gaze direction. After calibration, the gaze direction can be derived from the eye vector.

For use in future home television applications, calibration should be very easy. We have approached this issue by a special method using more than one light source. The accuracy depends on the number of light sources used. On the other hand, the segmentation of the pupil and the highlights (reflections) becomes increasingly difficult with a larger number of light sources, because the pupil would be largely covered by the highlights and there would be more reflections on eye glasses worn by a user (Fig. 18). Three to five light sources seem to be optimal.

Our self-calibrating gaze tracker requires initialization before it can be used for the first time [66]. In this phase, a reference eye is defined by calibrating the system with a reference user. The reference user will have to fixate a set of calibration points shown on the display and the corresponding gaze vectors will be registered. Then, a mapping function is calculated which maps the gaze vectors of the reference eye onto the calibration points. This way, both the relationship between the gaze vectors and the gaze direction, and the geometric arrangement of the camera and the light sources are taken into account in the calibration process. When a new user looks at the display, he/she will be presented an eye catching visual object instinctively attracting his/her gaze. At that moment, several gaze vectors are determined simultaneously. A user mapping function is deduced which maps the set of gaze vectors of the actual user onto the set of gaze vectors of the reference user, so that the actual user's gaze direction can be calculated.

In order to ensure exact measurements at any head position, we have introduced a head-fixed coordinate transformation that compensates for measurement deviations caused by head movements [67]. The gaze tracker runs on an ordinary PC with an update rate of 50 Hz and has an accuracy of about 0.7° after precise individual calibration and of about 1.2° when using the self-calibrating method, respectively. Fig. 19 shows that the algorithm is quite robust against distracting reflections on the user's spectacles. As long as these reflections do not seriously cover the pupil border, the pupil position and the corneal reflections can be correctly detected. We are currently implementing an extended version of the tracker, where selected LED sources can be activated and switched off, depending on whether or not they produce distracting reflections on spectacles.

C. Hand Gestures and Finger Tracking

Gesture input can be categorized into deictic gestures (point at an object or to a direction), mimetic gestures (accept or refuse an action), iconic gestures (define an object or its feature), and manipulative gestures (change the size and form of an object). From a technical viewpoint, the human hand is a composite nonrigid object. In order to recognize hand gestures, the position, orientation, and motion of significant limbs, such as the fingers and the palm of the hand, need to be detected and identified using user-worn devices (data gloves) or video-based contact-free techniques. Several researchers tried to model the human hand with a very high degree of freedom (about 26°) [68] or with hidden Markov models [69]. This way, it is possible to track the position and movement of the hand and to recognize even very detailed gestures, such as American Sign Language [70].

The Fraunhofer HHI hand tracker (Fig. 20) uses two infrared sensitive cameras in a stereo setup and an infrared light source next to the cameras. The infrared light is reflected by the hand and provides a bright image, since the hand is close to the light source. The intensity of light reflected from the distant background is comparatively weak. This simplifies the hand detection and segmentation process. During initialization, a reference image without the user's hand is captured. When the hand appears in the video image, it is detected by calculating the luminance difference between the current image and the stored reference image.

Luminance differences exceeding a predefined threshold indicate candidate regions of the hand in the image. The difference images undergo a temporal low-pass filtering to reduce noise effects. After a smoothing process using a morphological filter (for connecting small areas and further noise

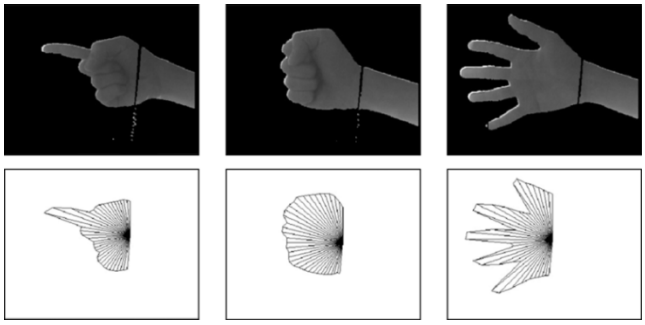


Fig. 21. Segmented hand regions and hand vectors indicating simple hand gestures.

reduction), the largest connected area defines the hand region in the current image. Next, a circle is matched to the palm area and a set of vectors pointing from the center of the circle to the boundary of the hand region (hand vectors) are calculated. Fingers are indicated by comparatively long vectors. Simple hand gestures are recognized by analyzing the distribution of the hand vectors and comparing it with prestored distribution patterns (Fig. 21). The hand segmentation and gesture recognition processes are performed on one of the two stereo images only. In a simplified approach the fingertips are found in the segmented images with a Hough-transform based filter; the front most fingertip is selected for touch interactions (Fig. 20).

A small set of feature points including the center of the palm and the finger tips are selected in the preprocessed image and their correspondences are searched for in the other stereo image using a feature point based block matching technique. Based on the geometrical parameters of the stereo camera setup, the 3-D position of the hand is calculated. Taking into account the position, gesture, and status of the hand (still or in motion) the interface software decides whether the user is currently moving or grasping an object or pointing at a target. Currently, the 3-D position of the fingertip is measured at an average accuracy of about 5 mm in the horizontal plane and about 15 mm in the vertical distance at 50/60 Hz update rate. The tracking range is 400 $\times$ 500 mm for a distance of 500 mm between the hand and the tracking module.

We achieve precise measurements in extended interaction spaces by cascading sets of the original tracking module, forming a multiple-baseline stereo tracker. Fig. 22 shows a prototype system, where the various display and multimodal interaction technologies were implemented in a video workplace. This system is, of course, not a visionary home TV set; we use it to develop and test novel ideas which may be used later in future interactive video applications.

V. Conclusion

"Being digital" is significantly changing our role in the information age. We are rapidly evolving from being passive consumers of media content toward becoming more and more (inter)active users. Thus, interactivity becomes the key concept of future media, including traditional services such as television. This paper outlines recent trends in a number

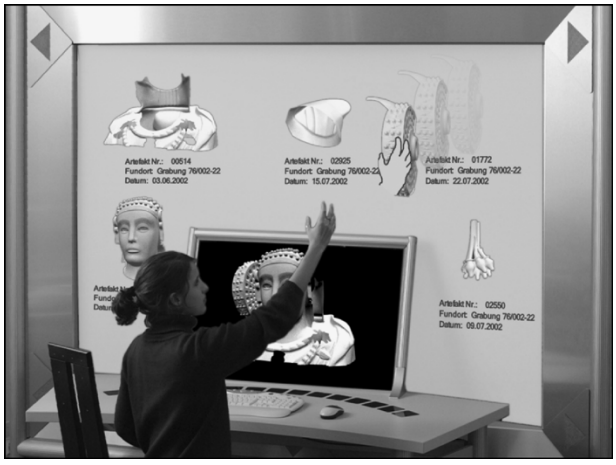


Fig. 22. Prototype video workplace allowing users to "drag" computer-generated or scanned 3-D video objects from a high-resolution rear-projection screen and "drop" them into the center 3-D display for direct manipulation. Cascaded video-based hand-gesture recognition devices are embedded in the desktop. Cameras behind the front plate of the 3-D display serve to recognize the viewer's gaze direction and head position. In connection with speech input, the computer vision devices are used for multimodal interaction.

of research areas, such as 3-D data representation formats, DIBR, autostereoscopic 3-D displays, and multimodel user interfaces. All together, these technologies could make the interactive 3-DTV of the future possible for single users (personal 3-DTV) as well as group viewing. The described state-of-the-art developments already show promising results in the labs; however, further development will be required to meet all 3-DTV user requirements.

Acknowledgment

The authors would like to thank the following individuals and institutions for the help providing data, results, and figures for this paper: K. Hopf, P. Kauff, F. Neumann, D. Przewozny, B. Quante, K. Schüür (Fraunhofer HHI, Germany), G. J. Iddan, G. Yahav (3DV Systems, Ltd., Israel), J. Mulligan, V. Isler, and K. Daniilidis (Penn State University, University Park).

References

[1] R. F. Tiltman, "How 'stereoscopic' television is shown," *Radio News,* 1928.
[2] DVB. 2005 [Online]. Available: http://www.dvb.org/
[3] P. Swan, *TV Dot Com: The Future of Interactive Television*. New York: TV Books, 2000.
[4] S. Curran, *Convergence Design: Creating the User Experience for Interactive Television Wireless and Broadband*. Gloucester, MA: Rockport, 2002.
[5] H.-Y. Shum and S. B. Kang, "A review of image-based rendering techniques," in *Proc. Visual Communications and Image Processing* 2000, pp. 2–13.
[6] K. Perlin, C. Poultney, J. S. Kollin, D. T. Kristjansson, and S. Paxia, "Recent advances in the NYU autostereoscopic display," *Proc. SPIE, Stereoscopic Displays and Virtual Reality Systems VIII* pp. 196–203, Jan. 2001.
[7] R. Bates, P. Surman, I. Sexton, M. Craven, K. C. Yow, and W. K. Lee, "Building an autostereoscopic multiple-viewer television display," in *Proc. Asian Symp. Information Display* 2004, pp. 400–404.
[8] W. Matusik and H. Pfister, "3D TV: a scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes," in *Proc. ACM SIGGRAPH* 2004, pp. 814–824.

[9] W. A. IJsselsteijn, P. J. H. Seuntiëns, and L. M. J. Meesters, State-of the-art in human factors and quality issues of stereoscopic broadcast television Eindhoven Univ. Technology, ATTEST Tech. Rep. D1, Aug. 2002.

[10] C. Wheatstone, "On some remarkable, and hitherto unobserved, phenomena of binocular vision," *Philos. Trans. R. Soc. Lond.,* 1838.

[11] A. Smolic and P. Kauff, "Interactive 3-D video representation and coding technologies," *Proc. IEEE,* vol. 93, no. 1, pp. 98–110, Jan. 2005.

[12] E. Catmull, "A subdivision algorithm for computer display of curved surfaces," Ph.D. dissertation, Univ. Utah, Salt Lake City, 1974.

[13] T. Matsuyama, X. Wu, T. Takai, and T. Wada, "Real-time dynamic 3-D object shape reconstruction and high-fidelity texture-mapping for 3-D video," *IEEE Trans. Circuits Syst. Video Technol.,* vol. 14, no. 3, pp. 357–369, Mar. 2004.

[14] C. Fehn, K. Hopf, and B. Quante, "Key technologies for an advanced 3D-TV system," *Proc. SPIE Three-Dimensional TV, Video, and Display III* pp. 66–80, Oct. 2004.

[15] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," *ACM Trans. Graph.,* vol. 23, no. 3, pp. 598–606, Aug. 2004.

[16] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. ACM SIGGRAPH* 1996, pp. 31–42.

[17] 3D Consortium. 2005 [Online]. Available: http://www.3dc.gr.jp/english/

[18] C. Fehn, P. Kauff, M. Op de Beeck, F. Ernst, W. A. IJsselsteijn, M. Pollefeys, L. Van Gool, E. Ofek, and I. Sexton, "An evolutionary and optimized approach on 3D-TV," in *Proc. Int. Broadcast Conf.* 2002, pp. 357–365.

[19] N. Hur, C.-H. Ahn, and C. Ahn, "Experimental service of 3DTV broadcasting relay in Korea," *Proc. SPIE, Three-Dimensional TV, Video, and Display* pp. 1–13, Jul. 2002.

[20] I. Yuyama and M. Okui, "Stereoscopic HDTV," in *Three-Dimensional Television, Video, and Display Technologies*. Berlin, Germany: Springer-Verlag, 2002, pp. 3–34.

[21] *Information Technology—Generic Coding of Moving Pictures and Audio: Video*, ISO/IEC 13818-2:1996, ISO/IEC JTC 1/SC 29/WG 11, Apr. 1996.

[22] *Information Technology—Coding of Audio-Visual Objects—Part 2: Visual*, ISO/IEC 14 496-2:1999, Dec. 1999, ISO/IEC JTC 1/SC 29/WG 11.

[23] *Information Technology—Coding of Audio-Visual Objects—Part 10: Advanced Video Coding*, ISO/IEC 14496-10:2003, ISO/IEC JTC 1/SC 29/WG 11, Jun. 2003.

[24] G. J. Iddan and G. Yahav, "3D imaging in the studio (and elsewhere . . .)," *Proc. SPIE, Videometrics and Optical Methods for 3-D Shape Measurement VII* pp. 48–55, Jan. 2001.

[25] M. Kawakita, T. Kurita, H. Kikuchi, and S. Inoue, "HDTV axi-vision camera," in *Proc. Int. Broadcast Conf.* 2002, pp. 397–404.

[26] R. I. Hartley and A. Zisserman*, Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[27] O. Faugeras, Q.-T. Luong, and T. Papadopoulo*, The Geometry of Multiple Images: The Laws That Govern the Formation of Multiple Images of a Scene and Some of Their Applications*. Cambridge, MA: MIT Press, 2001.

[28] J. Mulligan, V. Isler, and K. Daniilidis, "Trinocular stereo: a new algorithm and its evaluation," *Int. J. Comput. Vis.,* vol. 47, no. 1–3, pp. 51–61, Apr.–Jun. 2002.

[29] R. Szeliski and P. Torr, "Geometrically constrained structure from motion: points on planes," in *Proc. Eur. Workshop 3-D Structure from Multiple Images of Large-Scale Environments* 1998, pp. 171–186.

[30] M. Pollefeys, "3D modeling from images (tutorial)," presented at the Eur. Conf. Computer Vision, Dublin, Ireland, 2000.

[31] F. Ernst, "2D-to-3D conversion based on time-consistent segmentation," presented at the Proc. Workshop Immersive Communication and Broadcast Systems, Berlin, Germany, 2004.

[32] L. McMillan, "An image-based approach to three-dimensional computer graphics," Ph.D. dissertation, Univ. North Carolina, Chapel Hill, Apr. 1997.

[33] W. R. Mark, "Post-rendering 3-D image warping: Visibility, reconstruction, and performance for depth-image warping," Ph.D. dissertation, Univ. North Carolina, Chapel Hill, Apr. 1999.

[34] L. Lipton*, Foundations of the Stereoscopic Cinema—A Study in Depth*. New York: Van Nostrand Reinhold, 1982.

[35] A. Woods, T. Docherty, and R. Koch, "Image distortions in stereoscopic video systems," Proc. SPIE, Stereoscopic Displays and Applications IV pp. 36–48, Feb. 1993.

[36] G. Xu and Z. Zhang*, Epipolar Geometry in Stereo, Motion and Object Recognition*. Dordrecht, The Netherlands: Kluwer, 1996.

[37] P. Milgram and M. Krüger, "Adaptation effects in stereo due to on-line changes in camera configuration," *Proc. SPIE, Stereoscopic Displays and Applications* pp. 122–134, Feb. 1992.

[38] C. Fehn, "Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV," *Proc. SPIE Stereoscopic Displays and Virtual Reality Systems XI,* pp. 93–104, Jan. 2004.

[39] J. Shade, S. Gortler, L.-W. He, and R. Szeliski, "Layered depth images," in *Proc. ACM SIGGRAPH* 1998, pp. 231–242.

[40] Y.-Y. Chuang, B. Curless, D. H. Salesin, and R. Szeliski, "A Bayesian approach to digital matting," in *Proc. IEEE Computer Vision and Pattern Recognition* 2001, vol. II, pp. 264–271.

[41] *Information Technology—Generic Coding of Moving Pictures and Audio: Systems*, ISO/IEC 13818-1:1996, ISO/IEC JTC 1/SC 29/WG 11, Apr. 1996.

[42] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A transport protocol for real-time applications," RFC 3550, Feb. 2004.

[43] M. Wöpking, "Subjective assessments of 3-D versus 2-D presentation of motion pictures for large screen home television," presented at the Proc. 1st Int. Symp. 3-D Images, Paris, France, 1991.

[44] S. Pastoor, "Human factors of 3-D imaging: results of recent research at Heinrich-Hertz-Institut Berlin," in *Proc. 2nd Int. Display Workshop* 1995, pp. 69–72.

[45] N. A. Valyus*, Stereoscopy*. New York: Focal, 1966.

[46] T. Okoshi*, Three-Dimensional Imaging Techniques*. New York: Academic, 1976.

[47] S. Pastoor, "3D displays," in *3D Videocommunication*, O. Schreer, P. Kauff, and T. Sikora, Eds. Chichester, U.K.: Wiley, 2005.

[48] SeeReal Technologies 2005 [Online]. Available: http://www.seereal.com/

[49] D. Ezra, G. J. Woodgate, B. A. Omar, N. S. Holliman, J. Harrold, and L. S. Shapiro, "New autostereoscopic display system," *Proc. SPIE, Stereoscopic Displays and Virtual Reality Systems II,* pp. 31–40, Feb. 1995.

[50] G. Boerger and S. Pastoor, "Autostereoskopisches Bildwiedergabegerät (Autostereoscopic Display)," Germany, Patent 19537499, 2003.

[51] IRIS-3D, Ltd.. 2005 [Online]. Available: http://www.iris3d.com/

[52] P. Surman, I. Sexton, R. Bates, W. K. Lee, M. Craven, and K. C. Yow, "Beyond 3-D television: The multi-modal, multi-viewer, TV system of the future," in *Proc. 12th Int. Symp. Advanced Display Technologies (SID/SPIE FLOWERS)* 2003, pp. 208–210.

[53] C. van Berkel and J. A. Clarke, "Autostereoscopic display apparatus," U.S., Patent 6064424, 2000.

[54] S. Pastoor, "Research on 3-D imaging at Heinrich-Hertz-Institut Berlin," in *Proc. ITE Annual Convention, The Institute of Television Engineers of Japan* 1991, pp. 611–614.

[55] R. Börner, "Autostereoscopic 3D-imaging by front and rear projection and on flat panel displays," *Displays,* vol. 14, no. 1, pp. 39–46, 1993.

[56] ——, "Lenticages," Eur., Patent 0493863, 1998.

[57] S. Pastoor and K. Schenke, "Subjective assessments of the resolution of viewing directions in a multi-viewpoint 3-D TV system," *Proc. SID,* vol. 30, no. 3, pp. 217–223, 1991.

[58] T. Swedlow, "2000: Interactive enhanced television: A historical and critical perspective," in *Whitepaper Commissioned by the American Film Institute—Intel Enhanced Television Workshop* Jun. 2000.

[59] M. Uyttendaele, A. Criminisi, S. B. Kang, S. Winder, R. Szeliski, and R. Hartley, "Image-based interactive exploration of real-world environments," *IEEE Comput. Graph. Appl.,* vol. 24, no. 3, pp. 52–63, May/Jun. 2004.

[60] S. Pastoor and J. Liu, "3-D display and interaction technologies for desktop computing," in *Three-Dimensional Television, Video, and Display Technologies*, B. Javidi and F. Okano, Eds. Berlin, Germany: Springer-Verlag, 2002, pp. 315–356.

[61] N. Lesh, J. Marks, C. Rich, and C. L. Sidner, "Man-computer symbiosis revisited: achieving natural communication and collaboration with computers," *IEICE Trans. Inf. Syst.,* vol. E87-D, no. 6, pp. 1290–1298, 2004.

[62] faceLAB 2005 [Online]. Available: http://www.seeingmachines.com/
[63] A. Tomono, M. Iida, and K. Ohmura, "Eye tracking image pickup apparatus for separating noise from feature portions," U.S., Patent 5016282, 1991.
[64] B. Fröba and C. Küblbeck, "Face detection and tracking using edge orientation information," *Proc. SPIE, Visual Communications and Image Processing* pp. 583–594, Jan. 2001.
[65] L. Young and D. Sheena, "Methods & designs: survey of eye movement recording methods," *Behav. Res. Methods Instrum.,* vol. 7, no. 5, pp. 397–429, 1975.
[66] J. Liu and S. Pastoor, "Computer-aided video-based method for contactlessly determining the direction of view," Eur., Patent 1228414, 2003.
[67] J. Liu, "Device for determining a fixation point," U.S., Patent 6 553 281, 2003.
[68] V. I. Pavlovic, R. Sharma, and T. S. Huang, "Visual interpretation of hand gestures for human-computer interaction: a review," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 19, no. 7, pp. 677–695, Jul. 1997.
[69] D. Heckenberg and B. C. Lovell, "MIME: a gesture-driven computer interface," *Proc. SPIE Visual Communications and Image Processing,* pp. 261–268, Jun. 2000.
[70] T. Starner, J. Weaver, and A. Pentland, "Real-time american sign language recognition using desk and wearable computer based video," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 20, no. 12, pp. 1371–1375, Dec. 1998.

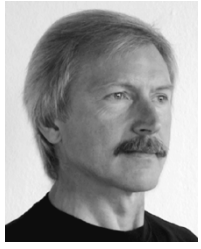
Christoph Fehn received the Dipl.-Ing. degree from the University of Dortmund, Germany, in 1998.

Since 1998, he has been with the Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institut (HHI), Berlin, Germany, where he is currently employed as a project manager in the image processing department. He has been involved in a number of national and international research projects, such as ACTS MoMuSys, ACTS CustomTV, DFG VITA, BMBF ITI, and IST ATTEST, where he conducted research in various fields of video processing, video coding and transmission, computer vision, and computer graphics. Most recently, he is working in the field of scalable video compression for digital cinema applications in the European IST project WorldScreen.

Mr. Fehn received the "President's Award" in 2001 for best technical paper at the International Broadcast Conference (IBC), Amsterdam, The Netherlands.

René de la Barré received the diploma degree in electronics technology and the Ph.D. degree from the University of Mittweida, Germany, in 1978 and 1993, respectively.

From 1978 to 2000, he worked in several positions in R&D for computer and visualization industries. Since 1994 he worked in the field of autostereoscopic multiview visualization for the VisuReal Displaysysteme GmbH. In 1998 he moved to the Innovation Centre of Plauen and led the region funded virtual prototyping project HOLOTRON. In 2001, he moved to Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institut (HHI), Berlin, Germany, and began his work on single-user 3-D display technologies in the national project multimo3D. Beginning in 2002, he became Subproject Leader for the HHI single-user 3-D display in the EU-funded 3-DTV project ATTEST. In the end of 2002, he became project leader for Project mixed3D. In this project, he was also responsible for research activities in human interaction technologies. He is currently Project Leader in the Interactive Media—Human Factors Department. His current research interests include 3-D displays and video based interaction technologies.



Siegmund Pastoor received the diploma and Ph.D. degrees in aerospace technology from the Technical University of Berlin, Germany, in 1975 and 1980, respectively. He was a lecturer of courses on man–machine systems from 1979 to 1983. of the Fraunhofer HHI. In 1980 he joined the Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institut (HHI), Berlin, where he is currently head of the Interactive Media—Human Factors Department. As a leader of a human-factors research group he has initiated and directed research activities in the fields of information display (character design, user guidance, use of color), 3-D imaging (psycho-optical foundations of 3-DTV, autostereoscopic 3-D displays, image processing for multiview 3-D systems) as well as in advanced display and interaction technologies for automotive applications. He was also a visiting scientist at NHK Labs, Tokyo, Japan, and has been involved in various European research activities. His current research interests focus on novel free-viewing 3-D display technologies and on nonconventional (multimodal) user interfaces for mixed reality applications.