1. Introduction

A video signal is a signal containing information for producing a visual image on a display device. The first such video signals, demonstrated in 1926, contained information on monochrome pictures of limited resolution and quality.

The video signal at the heart of the phenomenon known today as television is essentially very similar to the first one demonstrated in 1926. There are two essential differences. Colour information has been added to the signal. Also an audio signal, a signal containing information for producing sounds from a loudspeaker, is commonly transmitted along with the video signal.

In the realm of seeing and hearing the most fundamental enhancement which can be made to the experience of television is the addition of binocular depth perception. To achieve this appropriate image pairs must be displayed for each of the viewer's eyes. The information for the reconstruction of stereoscopic image pairs on a display device is contained in a stereoscopic video signal.

A brief overview of the components which contribute to the perception of depth is given in Section 1.1. Definitions of the commonly used term "stereoscopic" and "3-D" are given in Section 1.2. The reasons for adding stereoscopic information to the video signal are discussed in Section 1.3.

Given a stereoscopic video signal coding and compression of this signal will have to be appropriately handled. This is discussed in Section 1.4. To make the perception of depth as realistic as possible the images actually presented to the viewer should be adapted to the viewer's position relative to the display device. This operation is introduced in Section 1.5.

An overview of the contents of subsequent chapters of this thesis is given in Section 1.6.

1.1 Depth perception and depth cues

Depth perception is based on several factors. The factors which help the viewer to work out the relative distance of objects within a scene are referred to as *depth cues*. Many depth cues do not require the observer to have two eyes, they can be observed and appreciated just as well with one eye. These depth cues are knows as monocular or extrastereoscopic depth cues. There are a number of such monocular depth cues which are listed below. One very important depth cue is the binocular or stereoscopic depth cue which does depend on the observer having and using two eyes.

1.1.1 Monocular depth cues

Light and shade form an important depth cue [Valy62, Lipt91]. They are present in all imagery recorded from the real world, in photographs and in video material. The addition of light and shade to a hand drawn scene or a computer graphic enhances the feeling of depth. A fact which was already known to the artists of the Renaissance. Cast shadows provide a clue as to the relative position of objects. A simple scene with and without the addition of light and shade is shown in Figure 1.1.

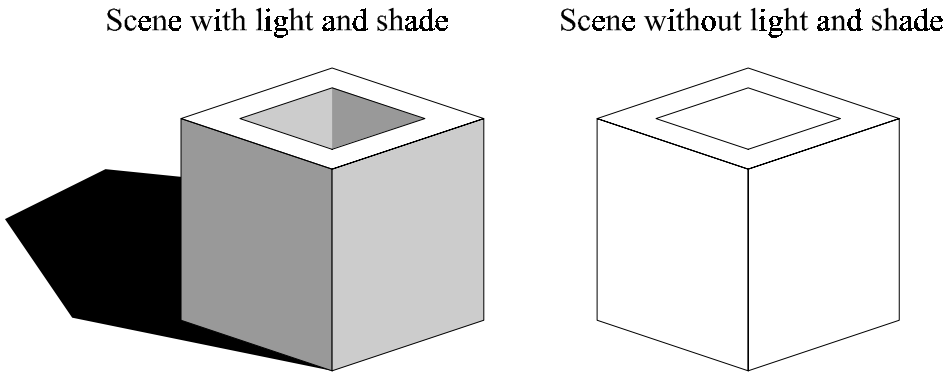

Figure 1.1: Light and shade as monocular depth cues.

Relative size is a depth cue [Valy62, Lipt91]. If two objects in a scene are known to have the same size because of a-priori knowledge held by the observer but they are

Figure 1.2: Relative size as a monocular depth cue.

located at different distances from this observer then the projection of the near object onto the observer's retina will be larger than the projection of the far object. In inspection of Figure 1.2 there is little ambiguity as to which of the two objects of interest is nearer, even in the absence of other depth cues.

Motion parallax [Valy62, Lipt91] is similar to the relative size cue: objects moving at a certain speed will appear to be moving faster when they are near to the observer than when they are far from the observer. When travelling in a train the landscape moves relative to the traveller. Trees near the railway track will pass by quickly while buildings which are near to the horizon only move slowly. Motion parallax works in a horizontal as well as in a vertical sense.

Interposition is the effect where an object which is located between the observer and a further away object (partially) occludes the far object [Valy62, Lipt91]. The occluding object must be nearer than the occluded objects. In Figure 1.3 the triangle must be nearer than the square but not as near as the circle which occludes them

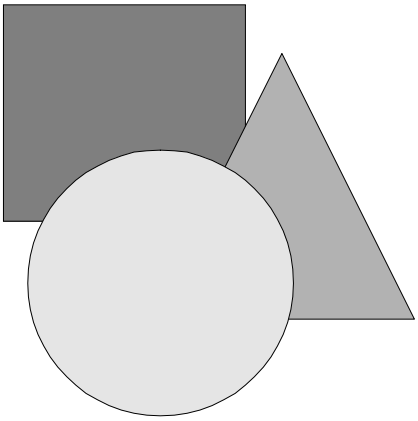

Figure 1.3: Interposition as a monocular depth cue.

both.

Textural gradient is a monocular depth cue because the texture on a textured material such as fabric or wood is more apparent when it is near than when it further away [Lipt91].

Aerial perspective is the reduction of visibility of distant objects because of haze or

Figure 1.4: Aerial perspective as a monocular depth cue.

fog. Distant scenes can get a bluish look because the atmosphere scatters red light while fog can quickly reduce visibility at distance [Valy62, Lipt91].

Geometric perspective is the effect where a (compound) objects appears to be getting gradually smaller as it gets further away from the observer [Valy62, Lipt91]. A typical example would be a corridor as illustrated in Figure 1.5. The dotted lines show how parallel straight lines appear to converge at a vanishing point.

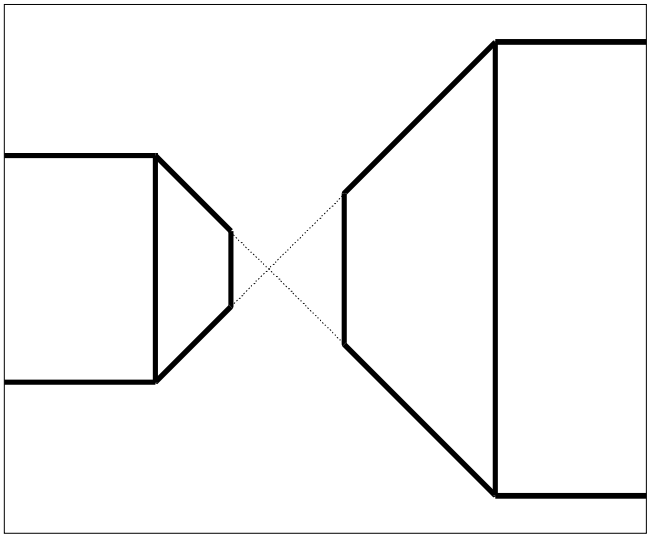

Figure 1.5: Geometric perspective as a monocular depth cue.

1.1.2 The binocular depth cue

The binocular depth cue relies upon the viewer having two eyes which perceive the scene which the viewer is watching. Each from a slightly different viewpoint so that

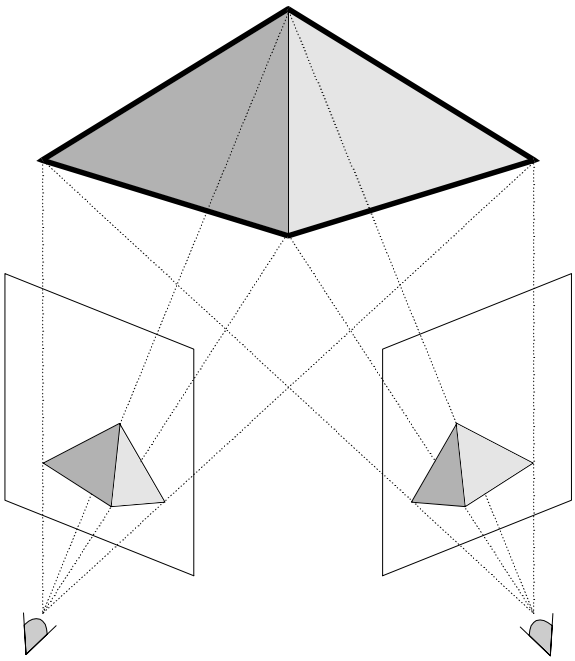

Figure 1.6: Left and right viewpoints of one scene.

slightly different images are projected on the respective left and right eye retinas (Figure 1.6). The brain then fuses the two to give depth perception, the viewer sees one solid scene instead of two slightly different projections. The perceived image with depth contains everything which is present in the two individual viewpoint images but adds what is present in neither of them: an intrinsic feeling of depth, distance and solidity.

The differences between the left and right eye viewpoint images lie in the fact that objects are relatively displaced according to distance from the viewer, according to depth. In digital images, where the function of the retina in the eye is taken over by a CCD in a camera, this displacement is referred to as *disparity*.

While the binocular depth cue is a necessary condition for full depth perception it is to note that it is by no means a sufficient condition. A stereoscopic scene with the binocular depth cue but without the other depth cues does not give a natural depth perception. If the other depth cues are inconsistent with the binocular depth cue then perceptual confusion and discomfort of the viewer will result. The use of conflicting depth cues for optical illusions is a well established technique. A well known creator of such illusions is M.C. Escher [Ernst78]. A sample of his work is given in Figure 1.7.

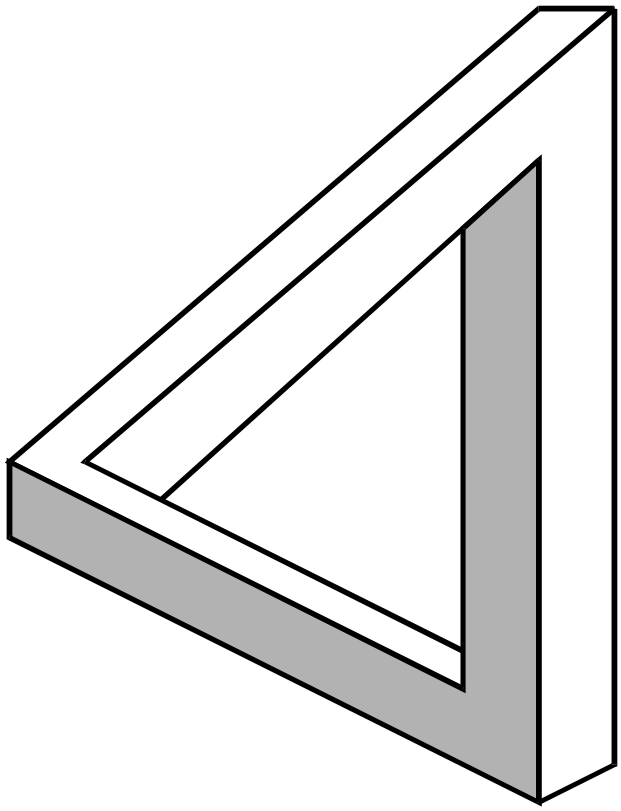

Figure 1.7: Confusing depth cues.

1.2 Stereo and 3-D

The terms stereoscopic and 3-D are often used interchangeably. This is not correct. In representing three dimensional solids and conveying depth there are three levels of realism.

- A single image which contains consistent and correct monocular depth cues is referred to as a stereographic image [Hawk91]. All imagery captured by a photographic or video camera is automatically stereographic. The scenery rendered by a flight simulation program on a monitor display is also stereographic: what is presented is the projection of a solid world onto the plane of the screen.
- Two stereographic images corresponding to left and right eye viewpoints (as illustrated in Figure 1.6) of a scene which are correctly presented to the left and right eye of a viewer form a stereo-pair or a stereoscopic image. If a rapid sequence of such images is presented then this is stereoscopic video.
- The limitation of stereoscopic images and stereoscopic video lies in the fact that the viewer is bound to watching the scene from the position of the camera: the viewpoint cannot be freely chosen. Systems where this viewpoint can be (up to a degree) freely chosen are called 3-D systems.

All 3-D images are stereoscopic and all stereo-pairs are composed of two stereographic images.

1.3 Reasons for stereoscopic video

Stereoscopic video adds to the realism of the viewer's television experience. This realism adds to the enjoyment of video in entertainment applications and it also instrumental in increasing the situational awareness of operators in (industrial) remote handling applications.

In the recent past it has been attempted to upgrade the resolution (the number of pixels) of the analogue domestic television signal in order to allow for bigger screens which cover a larger part of the viewer's field of vision. These attempts met with little success because of commercial reasons. The added realism did not weigh up against the costs of TV infrastructure replacement.

Well composed stereoscopy adds much more to visual realism than increased screen resolution. Solid objects are more realistic than plane projections. Also when the nature of the domestic television signal changes it will become digital. In this transition there lies the opportunity to add stereoscopic information right at the birth of a new television standard.

In remote handling applications machines are used to perform tasks in environments which are inaccessible to or too hostile for humans. Examples are sewers, access

pipes or sites which suffer from chemical or nuclear contamination. In the field of medicine one could think of endoscopic surgery.

These machines are remotely controlled by human operators. The operator sees the machine's working environment by means of a camera, mounted on or near the machine, which relays an image to a screen in front of the operator. Making this camera/screen combination stereoscopic greatly enhances the situational awareness of the operator. This allows the operator to more quickly adapt to new situations.

Monoscopic camera/screen configurations left some ambiguity as to depth. Monocular depth cues can, at best, convey relative depth. This means that hard tasks such as drilling require a period of practice when using a monoscopic camera/display because the operator has to learn to judge the depth in the scene. With a stereoscopic system the sensation of depth and solidity is absolute. The operator can go to work straight away. This produces a saving in time. Also it is valuable in applications such as surgery where the scope for errors on the first attempt is small or non-existent.

1.4 Stereoscopic video coding

With the advent of digital video signals has come the necessity for data compression of these signals because of the very large (relative to channel capacity) amounts of data involved. Digital video compression techniques operate by utilising redundancy in the digital video signal, discarding visually irrelevant information and encoding the remaining information as efficiently as possible.

Existing digital video compression techniques utilise redundancy in time by performing motion-based prediction. This is possible because subsequent images from a video sequence tend to be very similar. In stereoscopic digital video signals there is an additional kind of redundancy which may be exploited: left and right view images are also similar. Disparity-based prediction is an added possibility. Because of this extra redundancy it is possible to encode a stereoscopic digital video signal more compactly than two separate and independent monoscopic digital video signals.

In order for disparity-based prediction to be performed the disparity will first have to be estimated from the left and right views just as motion has to be estimated between subsequent images in conventional monoscopic codecs.

1.5 Intermediate viewpoint generation

A stereoscopic video system is not a 3-D video system. As the viewer moves the stereoscopic scene presented to him remains the same scene recorded by the stereoscopic camera which is probably not moving synchronously with the viewer's head. Looking "around" objects is not possible.

One solution to this problem would be to have a large number of camera's or a remote controlled mobile camera and to select a stereoscopic image pair to transmit and present to the viewer according to his position relative to the display. This may be practical in situations where there is a data channel available for each viewer, e.g. video conferencing, but in a broadcast context it would soon become impractical because a large number of different viewpoints would have to be simultaneously transmitted.

So assuming that a single stereoscopic video signal is available digital signal processing will have to be performed to synthesise intermediate views from this signal. This is referred to as intermediate viewpoint interpolation. This interpolation requires left and right view images. Information about the disparity of objects in these images will have to be estimated by a disparity estimator. In this application the demands on disparity estimation are very high because, unlike for coding, no additional information is available to compensate for errors in the synthesised images which are caused by faulty disparity values.

1.6 Disparity estimation

Both stereoscopic video coding and intermediate viewpoint generation require disparity vector fields. Coding techniques require them in order to make predictions while intermediate viewpoint generation requires them to link left and right view images in order to be able to generate intermediate viewpoints. Such a disparity field is illustrated in **Error! Reference source not found.**.

The exact nature of disparity will be dealt with in Chapter 2. In the disparity field illustrated in Figure 1.8 light grey-scale values correspond to low disparity (near to the viewer) while dark grey scale values correspond to high disparity (away from the viewer). This disparity vector field was calculated using full search block matching, a common technique for motion estimation.

Inspection of the figure will make it clear that this disparity field is not perfect.

- Some vectors are wrongly estimated and obviously bear no relation to the physical content of the scene.
- The vector field has been estimated using block based estimator. Further processing using this disparity field will show blocking effects.
- Some parts of the image displayed, the so-called occlusion areas, have no corresponding disparity. The full search block matcher does not take this into account.

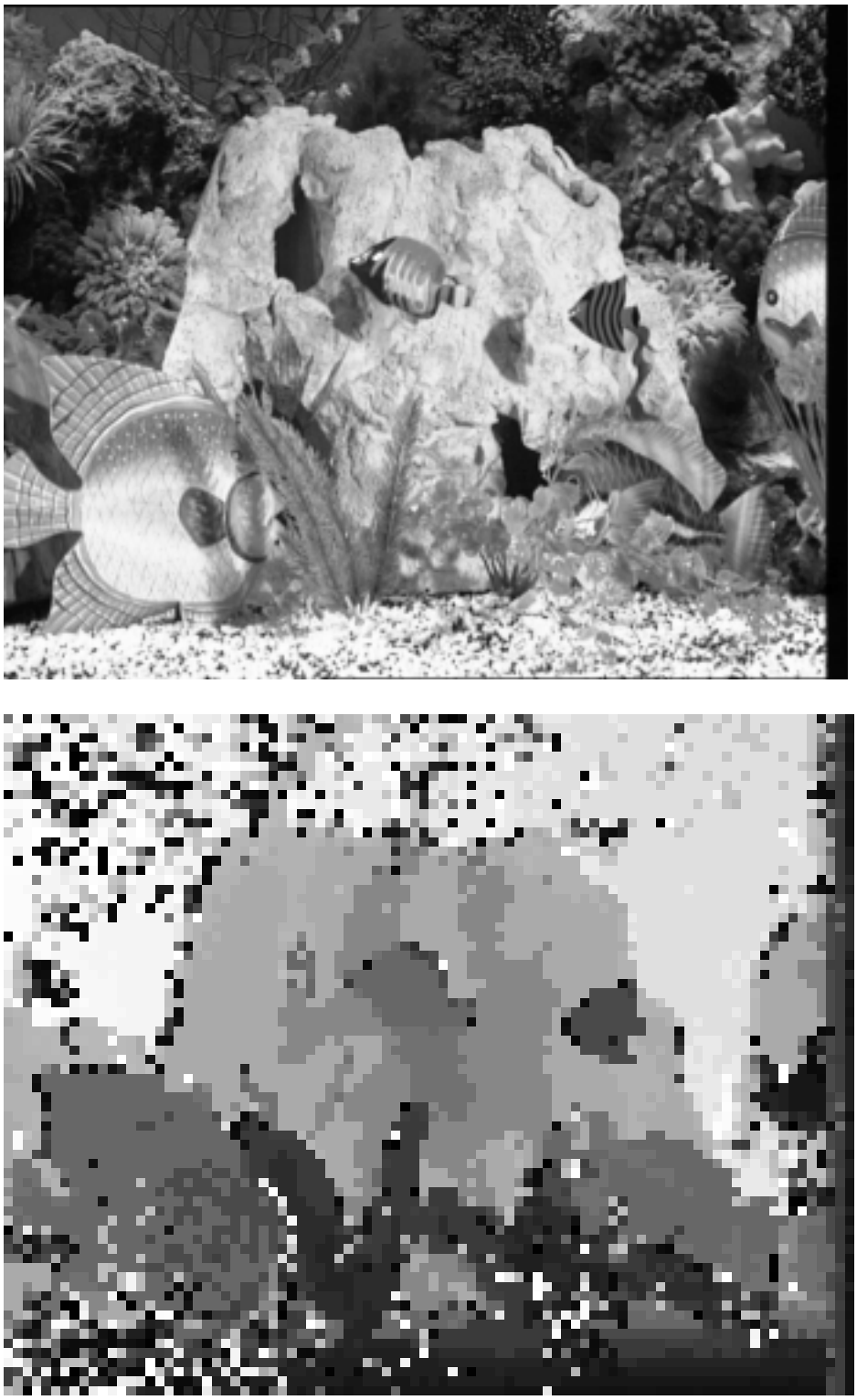

Figure 1.8: "Aqua" and full search estimated disparity field.

Disparity estimation for video coding is fairly robust against the problems mentioned above because the disparity field is used to calculate predicted images. These predicted images are then subtracted from original images and the resulting difference image is then encoded. Weaknesses in the disparity field are not immediately evident to the viewer.

Disparity estimation for intermediate viewpoint interpolation is a sensitive problem. Instead of just looking for a pixels which are similar we are looking for *the* pixel in the alternate image which is the projection of the same real world point which resulted in projection on the reference pixel(s). A disparity field like the one in **Error! Reference source not found.** contains too many locations where this correspondence does not hold. The block based nature of the disparity field means that the intermediate viewpoint images will also show blocking artefacts. Also the areas of occlusion are important to correct intermediate viewpoint generation. When the application is intermediate viewpoint generation estimating a disparity field which is almost right may very well produce results which are completely unacceptable.

1.7 The RACE-DISTIMA project

The RACE-DISTIMA (Research and technology development in Advanced Communications technologies in Europe - Digital Stereoscopic Imaging and Applications) project ran from 1992-1995 and concerned itself with a wide number of aspects of stereoscopic imaging and applications thereof.

As part of the project the stereoscopic chain of Figure 1.9 was realised in hardware and demonstrated.

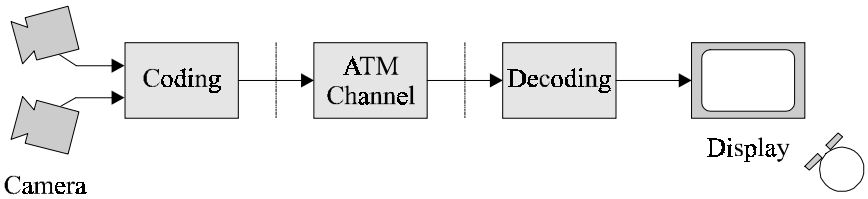

Figure 1.9: The DISTIMA stereoscopic chain.

The components of the chain were designed and produced by partners in the DISTIMA project.

The stereoscopic camera was built to appropriate mechanical specification so that precise alignment of the left and right view cameras is guaranteed. Also it's control mechanism was designed to automatically adjust those stereo-specific parameters which a camera operator would not normally adjust. The coding algorithm used was the MPEG-2 algorithm adapted to utilise the extra redundancy present in the stereoscopic video signal. The coding / decoding hardware was linked to an ATM network. After decoding at receiver side the stereoscopic video material is displayed on a stereoscopic projection system which was also developed within the DISTIMA project.

In order to design and make proper use of the video chain research was also done into perceptual aspects of stereoscopic video. Also several (industrial) applications of stereoscopic video were investigated.

Overall the DISTIMA project demonstrated the feasibility as well as the added value of stereoscopic video. It's successor project PANORAMA goes beyond the ground covered by DISTIMA by thoroughly investigating tele-presence applications as well as the various aspects of multi-viewpoint systems.

1.8 Overview of this thesis

In Chapter 2 the geometry of a stereoscopic camera configuration is discussed. Some basic results on the nature of disparity are derived. Luminance balancing to compensate for differences in camera electronics is discussed. Also a short description is given of the how and why of the hybrid motion/disparity/DCT codec which was implemented as a part of the DISTIMA hardware chain.

In Chapter 3 a brief introduction to existing blockmatching techniques is given. These techniques were originally developed for motion estimation. These techniques are evaluated in the context of disparity estimation. Some points of consideration for the design of a better disparity estimator are given.

The genetic vector field estimator discussed in Chapter 4 is an improved block matching technique. The genetic vector field estimator is a hybrid of conventional vector field estimators and the genetic algorithm. It estimates accurate and smooth vector fields without introducing undesirable constraints on the solution.

Disparity space image based techniques are dealt with in Chapter 5. The added value of these techniques lies in the fact that they can work at pixel resolution and that occlusion effects can be specifically modelled and handled.

Intermediate viewpoint generation, discussed in Chapter 6, is an uncompromising test of any disparity estimator. The actual functioning of intermediate viewpoint generation, given a stereo image pair and a disparity field, is described. All disparity estimators discussed in Chapters 3,4 and 5 are subsequently tested by using the disparity fields that they produce for intermediate viewpoint generation.

Fractal coding, which is the subject of Chapter 7, is a relatively new technique for image coding. Because it is a computationally very intensive technique it's applicability to video material (as opposed to still images) is not immediately obvious. In order to make fractal coding feasible for image sequences it is possible to use motion compensation. Not to make predicted images but to accelerate the search for self similar image parts.