# Joint disparity/motion estimation and segmentation for object-oriented stereoscopic image coding

# Joint disparity/motion estimation and segmentation for object-oriented stereoscopic image coding

*Cheng Hong Yang*

Université du Québec
**Institut national de la recherche scientifique**
INRS-Télécommunications
16, place du Commerce, Verdun
Québec, Canada, H3E 1H6

April 1997

Rapport technique de l'INRS-Télécommunications no. 97-05

This report describes the Ph.D thesis proposal of Cheng Hong Yang prepared under the supervision of Prof. Janusz Konrad and defended on April 18 1997 in front of a committee consisting of Prof. Benoît Champagne, Prof. Eric Dubois and Prof. Janusz Konrad.

# Summary

This document describes a project aiming at the development of a new approach to video segmentation. In particular, in order to divide individual video frames into regions corresponding to objects from a 3-D scene, the proposed approach attempts to jointly exploit multiple sources of visual information, such as motion, disparity and luminance/color. The resulting segmentation is expected to find applications in object-oriented coding of monoscopic and stereoscopic video sequences. The proposed approach is based on Markov random field models linked together by the maximum *a posteriori* probability estimation criterion. In the formulation robust error functions are used to minimize the impact of outliers, while the optimization stage is based on a continuation method. Encouraging initial experimental results for the case of motion segmentation are demonstrated.

# Contents

# List of Figures

# 1 Introduction

This project aims at developing a new approach to video segmentation. In particular, in order to divide individual video frames into regions corresponding to objects from a 3-D scene, it attempts to jointly exploit multiple sources of visual information, such as motion, disparity, luminance and color. The resulting segmentation is expected to find applications in object-oriented coding of monoscopic and stereoscopic video sequences.

The currently emerging moving picture standard MPEG-4 will provide many new functionalities that are not supported by the existing standards at low bit rates (MPEG-1, MPEG-2), such as content-based interactivity and bitstream editing [24]. In the context of MPEG-4, there is an increasing interest in object-based (segmentation-based) video coding. This interest is due to the fact that object-based approaches have a strong potential for increasing the coding efficiency as well as the inherent capability for handling content-based functionalities.

In the older moving picture coding standards interframe compression is achieved by motion compensation on rectangular blocks (*block-based*), hence at low data rates suffers from coding artifacts known as *blocking* and *mosquito* effects. In the framework of new standards a great attention is paid to object-oriented analysis-synthesis coding (OOASC). Instead of dividing images into blocks, they are divided into *moving objects* based on the assumption that an object corresponds to a region with uniform motion and luminance/chrominance characteristics. Then, motion compensation can be applied to every object, and each object can be coded by three sets of parameters defining its motion, shape and luminance/color information. It should be clear that the success of motion estimation is closely related to the accuracy of motion segmentation. A better quality of prediction should be achieved by object-oriented processing and hence higher compression ratios can be expected.

As far as stereoscopic video is concerned, the scenario of disparity estimation and compensation on left and right images is similar to that of motion estimation and compensation. Object-based techniques are also desirable in expectation of higher compression gains. Besides, we hope to obtain a more precise description of objects in the sequence because the additional camera provides depth information that we couldn't have in the case of a monocular system. If a good segmentation is available, every 3-D object can be constructed based on the depth map, and hence the reconstruction of an image from an arbitrary view point becomes a simple synthesis process [26]. Providing "look-around" capability in multi-view environments is also an extension of MPEG-4. Nevertheless it's on the condition that a good segmentation is available. Roughly speaking, the intermediate view reconstruction can described as an interpolation problem based on images from the existing view points. Therefore, it is important to ensure that the corresponding points from different images belong to the same object.

To achieve high quality for a human observer a good segmentation has to take luminance and color information into account. A segmentation which coincides with object shapes is more stable temporally, although it has been shown that coding an individual image based on rate-distortion theory does not require such a meaningful segmentation [45]. A temporally stable segmentation can be predicted based on motion compensation; its temporal redundancy is reduced. Furthermore, such a segmentation helps the coder to adapt its coding strategy to human visual system's properties.

The present project focuses on simultaneous multiple-source segmentation and estimation of, e.g., motion, disparity. We will develop a Markov random field (MRF) approach to model the relationship between the field of segmentation labels and individual sources. A MRF approach is very attractive as it can simultaneously take all elements into account rather than treating them one after another as most approaches do. Within our framework, we will also exploit a robust estimator in order to efficiently deal with statistical outliers.

# 2 Literature review

In this section we present some elementary concepts involved in our project. We begin with presenting Markov random fields (MRFS) and Gibbs random fields (GRFs). The original work of an object-oriented analysis-synthesis coding (OOASC) is presented in Section 2.2. In Section 2.3 we give a brief description of a stereo system and related problems. Motion segmentation techniques are reviewed in Section 2.4. In Section 2.5 we describe the basic principles of highest confidence first and continuation optimization methods. The concepts of robust estimators are reviewed in Section 2.6.

## 2.1 MRFs and GRFs

A random field $Z = \{Z(\mathbf{x}), \mathbf{x} \in \Lambda\}$ is a stochastic process defined over a lattice $\Lambda$. In our context $\Lambda$ belongs to the image plane, $\Lambda \subset \mathcal{R}^2$. Sites $\mathbf{x} = (X, Y)$ in $\Lambda$ correspond to positions of pixels. A random field $Z$ can be discrete-valued or real-valued. A *neighborhood* of a site $\mathbf{x}$ is a set of sites $\mathcal{N}(\mathbf{x})$ which has the properties:

- $\mathbf{x} \notin \mathcal{N}(\mathbf{x})$,
- $\mathbf{y} \in \mathcal{N}(\mathbf{x}) \Longleftrightarrow \mathbf{x} \in \mathcal{N}(\mathbf{y})$.

A neighborhood system $\mathcal{N}$ over $\Lambda$ is the collection of neighborhood of all sites.

Markov random fields are extensions of Markov chains to 2D lattices. The random field $Z$ is called a MRF with respect to $\Lambda$ if

$$p(Z(\mathbf{x})|Z(\mathbf{y}), \forall \mathbf{y} \neq \mathbf{x}) = p(Z(\mathbf{x})|Z(\mathbf{y}), \mathbf{y} \in \mathcal{N}(\mathbf{x})).$$

The definition of a Gibbs distribution is closely related to a structure in $\Lambda$ called a *clique.* A clique $c$ on $\Lambda$ with respect to the neighborhood system $\mathcal{N}$ is a subset of $\Lambda$ such that either $c$ consists of a single site or all pairs of sites in $c$ are neighbors. The set of all cliques is denoted by $\mathcal{C}$.

The Gibbs distribution, with neighborhood system $\mathcal{N}$ and the associated set of cliques, is defined as

$$p(Z = z) = \frac{1}{S} \exp(-\mathcal{U}(z)),$$

where

$$\mathcal{U}(z) = \sum_{c \in \mathcal{C}} V_c(z),$$

and $\mathcal{U}(z)$ is called the energy function. $V_c(z)$ is a *potential* associated with clique $c$, while

$$S = \sum_{z} \exp\left(-\mathcal{U}(z)\right),$$

called the partition function, is a normalizing constant. The only condition on the clique potential, otherwise totally arbitrary, is that it depends only on pixel values in cliques $c$. The above expression has the physical interpretation that the smaller $U(z)$ (the energy of the realization $z$), the more likely that realization (larger $p(Z = z)$).

The Gibbs distribution is basically an exponential distribution. The origins of GRFs lie in physics and statistical mechanics literature. The exponent is frequently expressed as $-\frac{1}{T}U'(z)$, where $T$ is called the "temperature", and this distribution is used in optimization by simulated annealing (see [25], [32]).

The valuable Hammersley-Clifford theorem, which provides us with a simple and practical way to specify MRFs through Gibbs potentials, proves the equivalence between a MRF and a Gibbs distribution on the same neighborhood system $\mathcal{N}$. Consequently, the energy function of a Gibbs distribution is a more convenient and natural mechanism for embodying image attributes than are the local characteristics of a MRF. The theorem brought a large number of MRF applications in image processing, e.g., image segmentation, motion, disparity and occlusion estimation.

A frequently used criterion in MRF-based estimation is the maximum *a posteriori* probability (MAP) criterion. Let $\mathbf{Z}$ be a random field of attributes to be estimated, e.g., motion and disparity, and $\mathbf{O}$ be a field of observation variables. The optimum $\mathbf{z}^*$ based on realization $\mathbf{o}$ can be found via the following MAP optimization:

$$\mathbf{z}^* \;=\; \arg\max_{\mathbf{z}} P(\mathbf{Z} = \mathbf{z}|\mathbf{O} = \mathbf{o})$$

$$\begin{aligned} &= \arg\max_{\mathbf{z}} \frac{P(\mathbf{O}=\mathbf{o}|\mathbf{Z}=\mathbf{z})P(\mathbf{Z}=\mathbf{z})}{P(\mathbf{O}=\mathbf{o})} \\ &= \arg\max_{\mathbf{z}} P(\mathbf{O}=\mathbf{o}|\mathbf{Z}=\mathbf{z})P(\mathbf{Z}=\mathbf{z}). \end{aligned}$$

According to the Hammersley-Clifford theorem, the problem is equivalent to the minimization of the energy function of the Gibbs distribution

$$\mathbf{z}^* = \arg\min_{\mathbf{z}} \mathcal{U}(\mathbf{O}=\mathbf{o}|\mathbf{Z}=\mathbf{z}) + \mathcal{U}(\mathbf{Z}=\mathbf{z}). \tag{1}$$

The first term is related to the likelihood model while the second term is related to the *a priori* model.

## 2.2 Object-based coding

### 2.2.1 Basic idea

To know what is object-oriented coding, we have to present the original work of Musmann, Hotter and Ostermann [22, 21, 33, 35, 36].

In [34], they described and implemented a complete object-oriented analysis-synthesis coder. An object-based analysis-synthesis encoder can be characterized by these special steps [47]:

1. *image analysis*: The frame to be encoded is segmented into individually moving objects using information from the previously encoded frames. Each object in the frame is characterized by a set of parameters (a model).

2. *image synthesis*: The present frame is synthesized based on the estimated attribute parameters and the information from previously encoded frames.

3. *coding*: The parameters are encoded by suitable coding methods.

In an object-oriented coder a more efficient compression should be possible based on the fact that every region contains many blocks that can be described by only a couple of parameters and that within every region the correlation is high. Moreover the human visual system is less sensitive to errors in the presence of high luminance contrast. Experimental results to date show that object-oriented coders are able to synthesize images which look more natural than images predicted by a block-based hybrid coder, although the mean square prediction error may be the same or even higher [34].

In an object-oriented coder in addition to motion information, the shape of regions must be transmitted; this shape limits the area to which a particular set of motion parameters applies. At the same time, since motion-based prediction is usually far from perfect, luminance/color residual (error) information needs to be transmitted as well. Therefore, the problem of coding the shape and luminance/color is important.

In conclusion, compared to block-based coding, object-based coding has to deal with three important problems:

1. Motion segmentation: we have to extract moving objects from the images. This is the core of our project; we will return to this problem in Section 2.4.

2. Luminance/color coding,

3. Shape coding.

### 2.2.2 Luminance/color coding and shape coding

For luminance/color coding a representative approach is that of Gilge *et al.* [14]. DCT was extended to a more general form: a linear combination of orthogonal basis functions defined on image region with arbitrary shape [14]. For each region an orthogonalization of the transform basis functions has to be performed separately using Gram-Schmidt algorithm.

The main disadvantage of this approach is the computational effort which is necessary for the orthogonalization of the basis functions. For each object shape an orthogonalization of the transform basis functions has to be performed separately, the computational load depends on the number of pixels within the object. A simpler coding scheme which essentially uses the block-based DCT and takes into account the segment information of the object, has been proposed by Schiller and Hotter [40]. This approach has a simpler computation scheme with respect to the method proposed by Gilge (generalized DCT). Schiller and Hotter have shown that the performance of the two methods is similar and much better than that of block-based DCT.

The shape of an object can be described either exactly or approximately. Lossless encoding the contour is usually more expensive than coding with respect to a fidelity criterion; it usually requires about 1.2 bits/contour-point [14]

In approximation methods, the distance between the original and approximated shapes is an important measure of approximation quality. A maximum error of two pixels horizontally, vertically and diagonally to the outside of a model-compliant object and one pixel to its inside has been found to keep the synthesis errors still acceptable [13, 18]. A polygonal representation of object shape is interesting for its simplicity, but it does not necessarily provide a natural-looking shape. In [21], Hotter improved polygonal representation by a combination of polygon and spline representation. First, the shape is approximated by a polygonal representation. The quality is controlled by the absolute distance between the approximate and the original object shape. By adding new vertices the polygon representation can be forced to meet a quality criterion. In the final step, polygon vertices are used to compute a spline approximation of the shape. If this representation satisfies the quality criterion, the

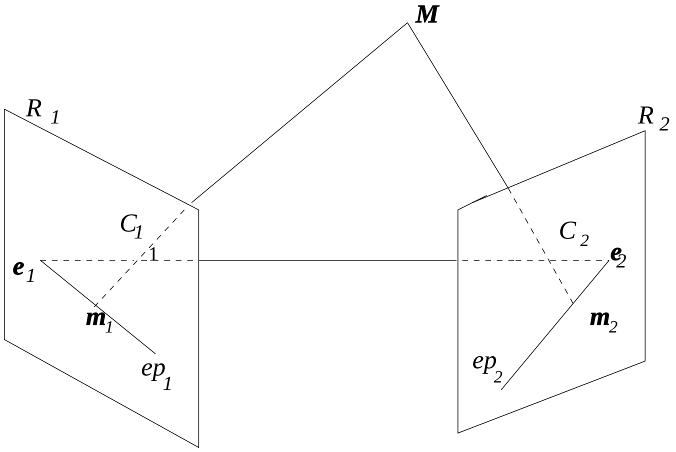

Figure 1: Epipolar constraint: a 3-D point $\mathbf{M}$ is projected onto two images planes ($R_1$ and $R_2$) through the optical centers $C_1$ and $C_2$.

spline representation replaces the polygon representation.

## 2.3 Stereoscopic video

### 2.3.1 Stereoscopic video system

Our eyes see the world from different points of view. Two slightly different images are obtained on our left and right retinas. The mind combines two different, although similar, images into one image (*fusion*) and the depth can be perceived due to the difference between the images (*disparity*).

A stereoscopic vision system consists of two or more cameras (Figure 1). To reconstruct 3D coordinates from a pair (or more) of given 2D images obtained by the cameras, we must first deal with the *correspondence problem*: given a token in image 1 what is the corresponding token in image 2 (see Figure 1). Because there are too many potential candidates to choose, some properties must be exploited. The fundamental constraint is the *epipolar* constraint.

In Figure 1, the point $\mathbf{e}_2$ ($\mathbf{e}_1$), which is the intersection point of the line $C_1C_2$ (*base line*) and the plane $R_2$ ($R_1$), is called the epipole of the second (first) camera with respect to the first (second) camera. The plane $C_1MC_2$ is called the *epipolar plane* and is defined by $\mathbf{M}$. We reason that all possible matches $\mathbf{m_2}$ in $R_2$ of $\mathbf{m_1}$ must be located on the images of the half line $ep_2$ (*epipolar line*). The epipolar lines are intersections of the epipolar plane and the image planes $R_1$ and $R_2$ [11].

While a stereoscopic system consists of a pair of video cameras mounted side by side, a stereoscopic display consists of a single display surface on which the left and right images are displayed and separated by a suitable method. The geometry of a

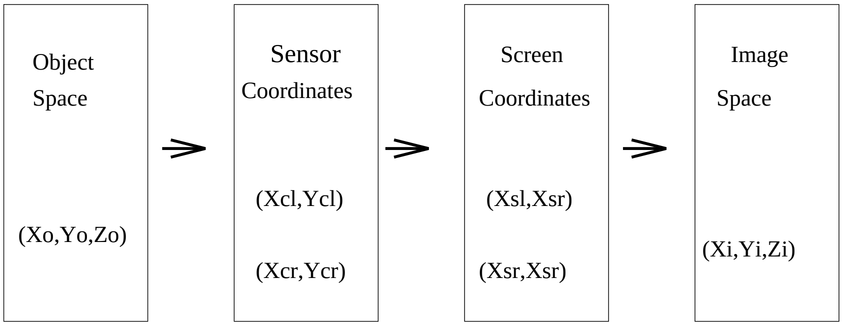

Figure 2: Sequence of transformations between the true 3-D scene and viewer-perceived 3-D space.

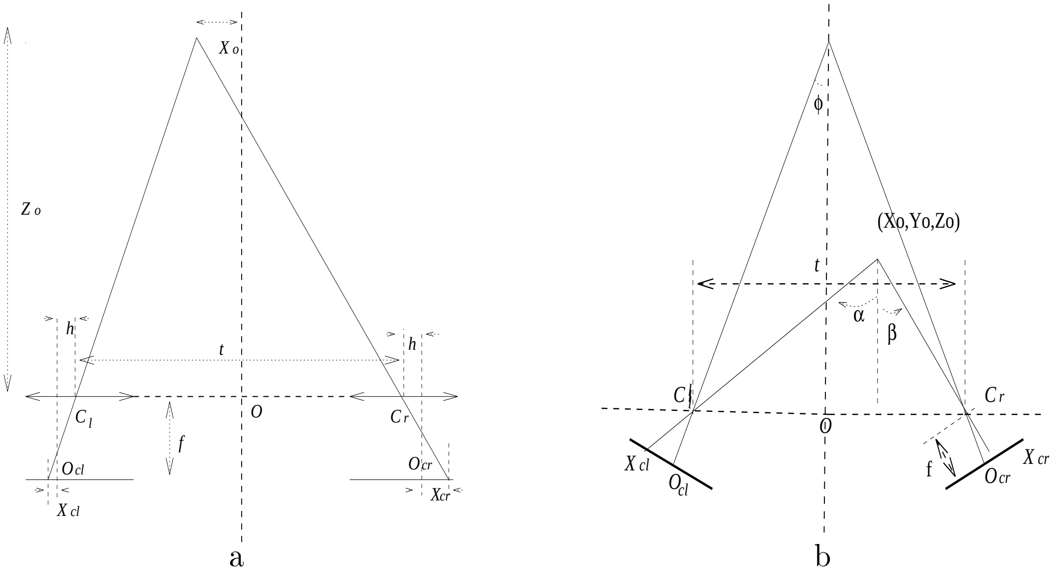

Figure 3: Parallel and convergent (toed-in) camera geometries.

stereoscopic video system can be determined by considering the imaging and display processes as three separate coordinate transformations (Figure 2). The first transformation is from 3D coordinates of the object in the real world (object space) to the position on the two sensors. The second transformation, is from these coordinates to the coordinates on the display screen, and the third one to a 3D position perceived by the viewer.

To develop the first transformation, camera geometry has to be considered. There are two types of camera setups with respect to the convergence: parallel cameras and toed-in or convergent cameras [5, 19].

The coordinates on the two sensors are given by Woods, Docherty and Koch [56] (refer to Figure 3). For parallel cameras, we have

$$\begin{aligned} X_{cl} &= \frac{f(t+2X_o)}{2Z_o} - h, && (2) \\ X_{cr} &= -\frac{f(t-2X_o)}{2Z_o} + h && (3) \\ Y_{cl} &= Y_{cr} = \frac{Y_o f}{Z_o} && (4) \end{aligned}$$

For toed-in cameras, We have:

$$\begin{aligned} X_{cl} &= f \tan(\alpha - \phi), && (5) \\ X_{cr} &= -f \tan(\beta - \phi) && (6) \\ Y_{cl} &= \frac{f Y_o \cos\alpha}{Z_O \cos(\alpha - \phi)} && (7) \\ Y_{cr} &= \frac{f Y_o \cos\beta}{Z_O \cos(\beta - \phi)} && (8) \end{aligned}$$

where

$$\alpha = \arctan \frac{t+2X_o}{2Z_o}, \qquad \beta = \arctan \frac{t-2X_o}{2Z_o}.$$

The distance between the coordinates of homologous points in the left and right images is referred to as *disparity*. From the above expressions, it's clear that there is no vertical disparity for parallel cameras. Many computations are greatly simplified in this case.

The transformation from the sensor coordinates to the screen coordinates is simply a magnification process. We don't need, in the following development, the formulations of the third transformation which describes how the *parallax* produces a stereoscopic cue.

### 2.3.2 Disparity estimation

Disparity estimation from stereoscopic images is similar to motion estimation from moving images. Therefore, many motion estimation techniques, such as pixel-, block- or segmentation-based, have been adopted for disparity estimation. However, disparity is characterized by several distinguishing features that need to be taken into account while designing an estimation procedure:

1. disparity often reduces to a scalar (parallel geometry),

2. disparity is usually non-zero almost everywhere in the image (except for areas resulting from a projection of surface in the camera convergence plane), whereas motion is often zero in large parts of an image (background obtained by a stationary camera),

3. the dynamic range of disparity is usually much larger than that of motion,
4. in practice disparity does not obey view-angle continuity unlike motion's temporal continuity (too spare view-angle sampling).

I just present here an interesting conclusion by Tamtaoui and Labit [46] and some methods using MRF model.

The kinetic constraint describing the relationship between disparity vector and motion vectors is often used to improve robustness [20, 53]

$$\mathbf{v}_l - \mathbf{v}_r + \mathbf{d}_t - \mathbf{d}_{t+1} = 0, \qquad (9)$$

where $\mathbf{v}_l, \mathbf{v}_r$ are motion vectors, $\mathbf{d}^t, \mathbf{d}^{t+1}$ are disparity vectors at $t$ and $t+1$.

Tamtaoui and Labit [46] have tested and compared three approaches to disparity estimation that all used the kinetic constraint. The results showed that for a good quality reconstruction disparity and motion should be jointly estimated.

Woo and Ortega [54] modeled the disparity and occlusion fields using MRFs(but an element corresponds to a block of pixels rather than a pixel). The final expression contains three terms: observation (likelihood) model which corresponds to the compensation error (DFD); the smoothness constraint adapted to occlusions, and the discontinuity constraint which imposes an *a priori* assumption on the occlusion.

Following the same principle, they addressed disparity estimation/segmentation problem in [55]. By comparison with the previous work, an *a priori* assumption has been added and the smoothness constraint is active only within the regions. Unfortunately, optimization procedure beginning with a simple luminance based segmentation is also sequential, i.e., disparity estimation and segmentation steps are interleaved.

### 2.3.3 Stereoscopic video coding

Similarly to motion compensation, disparity compensation reduces spatial redundancy between left and right images. Perkins [37] established a random process model for a stereopair source, and showed, based on the Shannon's theorem, that a stereopair source can be optimally noiselessly encoded and decoded by a structure that employs the following strategy: encode right image, encode left image given encoded right image; decode right image, decode left image given decoded right image.

Many block-based approaches to stereo coding (mainly MPEG-2 compatible encoders/decoders) [38, 49, 28, 16], and to multiview stereo coding [51, 50] have been reported. Two object-oriented implementations of stereoscopic coding have been developed [15, 52].

The coder proposed by Grammalidis *et al.* [15] works as follows. A block-based matching is performed at first to estimate the disparity. The resulting disparity map is stored for further motion estimation. The 3D motion is described by six parameters (**R, T**) for each view. Motion is estimated by minimizing the sum of displaced frame differences of left and right sequences. A fall-back block-based mode ensures satisfactory performance even when object-based compensation fails.

The method proposed by Tzovaras *et al.* [52] is an attempt to describe an object as motion and disparity object; an object is a region with similar motion and depth parameters. In the first step, a split-and-merge segmentation scheme is applied to segment the image into regions with uniform motion described by 6 parameters (translation and rotation). Each region is divided into blocks with a pre-defined size. An additional merge procedure is applied to the blocks to form regions of constant depth, the so-called *depth regions.* Each depth-region is defined by the following equation:

$$z(x, y) = Ax + By + C.$$

Therefore, each object is described by a motion parameter set and a disparity parameter set.

Both approaches [15] [52] employed the split-and-merge process to perform a segmentation on the pre-estimated motion and depth fields. From the results reported, a strategy better than "estimate and then segment" is desirable to improve both segmentation and estimation. It's clear that estimation (motion and disparity) and segmentation should be treated simultaneously to attain better quality. This problem can be addressed by using a MRF model.

## 2.4 Segmentation

Object extraction means the process of finding a region in the image which corresponds to the projection of a 3-D moving object. In case of moving object, it's referred to as motion (motion-based) segmentation.

### 2.4.1 Motion models

Every moving object, in most cases, can be described by a motion model. Therefore, motion model is involved in a motion segmentation process. An object model provides a relationship between the position of pixel at $(X_t, Y_t)$ in the image $I^t$ and the position $(X_{t+1}, Y_{t+1})$ based on a camera model and a geometric description of object's surface.

A planar rigid object with 3D motion (translation and rotation) was first studied by Tsai and Huang [48]. An eight-parameter model was developed. Four models known as 2D rigid source model with 3D motion, 2D flexible source model with 2D motion, 3D rigid source model with 3D motion and 3D flexible model with 3D motion, were implemented and compared by Hotter [21] and Ostermann [35].

Many other models have been also proposed. A review of such models can be found in [41]. Furthermore, many experiments show that affine transformation is a good enough motion model for real video sequences. Not all object-based approaches describe every region by a motion model. For example in [43, 44] a dense field was used instead.

### 2.4.2 Clustering techniques

The simplest way to perform a motion segmentation is to apply clustering techniques to a pre-estimated motion field. Clustering methods classify a set of entities (pixels) into a number of subsets according to some dissimilarity/similarity criterion. For spatial segmentation the criterion is usually based on color, luminance or texture information.

Many spatial segmentation technique exist in computer vision. Some of them follow very intuitive reasoning. The split-and-merge approach is one that is frequently used because of its simplicity. The application of split-and-merge to motion or disparity segmentation is straightforward [15, 52]. It is performed in two stages: split different entities into different classes according to a split criterion (dissimilarity criterion) and then merge homogeneous entities to form a new class according to a fusion criterion (similarity criterion). In the particular case of a quadtree, we have:

1. in split step: every region satisfying the split criterion is divided into four identical regions,

2. in merge step: every two adjacent regions satisfying the fusion criterion are merged into a new region.

A derivative of the standard split-and-merge approach which operates on flexible triangular meshes has been developed as well [30].

Another segmentation technique is region growing. The central idea is to choose some initial data units as seeds and then assign every pixel to the seeds. The technique can be considered as a variation of the so called non-hierarchical clustering methods with number of clusters $k$ specified a priori or determined as part of the clustering methods. Among the family of nearest-centroid sorting [2] methods in clustering analysis, MacQueen's $k$-means method [31] is simple and frequently employed in image analysis. The algorithm sorting the data units into $k$ clusters is composed of the following steps:

1. take the first $k$ units as clusters with a single member,

2. assign each of the remaining units to the cluster with the nearest centroid; after each assignment, recompute the centroid of the gaining cluster,

3. after all data units have been assigned in step 2, take the existing cluster centroids as fixed seed points and make one more pass through the data set assigning each data unit to the nearest seed point.

MacQueen proposed a variation on his basic $k$-means method which permits the number of clusters to vary during the initial assignment of the data units to clusters.

### 2.4.3 Motion segmentation

A motion segmentation approach is employed in many object-based analysis-synthesis coding schemes, for example [22, 34, 21, 35]. Instead of performing the identification of moving objects and estimation of mapping parameters in two steps, they are combined in one procedure. The segmentation and mapping parameters are determined jointly.

1. First, by simply comparing two consecutive frames $\mathbf{I_t}, \mathbf{I_{t+1}}$ two types of regions are distinguished: changed regions and unchanged regions.

2. Each connected changed region is considered as a moving object. A motion model and a geometric description of the surface of the object are employed. Based on the established model (motion parameters and geometric parameters), a mapping is sought that assures a correspondence between the image $\mathbf{I}_{t+1}$ and a predicted image $\hat{\mathbf{I}}_{t+1}$. For each region a parameter set is determined. A criterion measures the accuracy of mapping parameter estimation: it compares $\hat{\mathbf{I}}_{\mathbf{t+1}}$ and $\mathbf{I}_{t+1}$. If the model describes the moving region well, the object is determined and so are the mapping parameters. If a threshold is exceeded, the region will be treated in the next hierarchical step.

3. Each region is divided into a region which can be described well by the model and a model failure region. The model failure subregion is sent to step 2. This scheme proceeds hierarchically until each object is characterized by a parameter set.

The drawback of this method is that it can correctly identify different moving objects only when there is a dominant moving object.

Following the same principle, the approach of Diehl [10] takes the color information into account. The assumption is that moving objects and the background can be distinguished by significant contours. Thus, the boundaries of changed regions should coincide with contours extracted from individual images. A contour detector is applied and the motion segmentation is refined using the contour information.

### 2.4.4 MRF approaches

Using a MRF model, the motion and segmentation fields can be integrated easily into one framework.

Konrad *et al.* [27] established a motion/segmentation model that is similar to that of Woo *et al.* [54] for disparity described in Section 2.3.2. They applied the same assumption about moving objects' contours described above and employed an intensity segmentation as initial state. The segmentation was then refined by region fusion and boundary adjustment.

In the motion/segmentation model established by Stiller and Hurtgen [42], the observation model, the counterpart of the compensation error distribution is provided be a zero-mean *generalized Gaussian distribution*. In a subsequent work [44], a more sophisticated a priori model contained four terms which favor spatially-smooth segmentation, temporally continuous segment boundaries along motion trajectories, spatially smooth motion field within each segment, and temporally continuous motion vectors along motion trajectories. From a guess of motion and segments, the optimization problem is solved by the ICM method. The segmentation result reported is of good quality.

Simultaneous motion estimation and segmentation has also been addressed by Chang *et al* [4]. The authors include the affine motion model and the dense field model in one MRF formulation, and they solve the optimization by interleaving a segmentation problem and a motion estimation problem. First, by fixing the segmentation, they solve the resulting sub-optimization problem for the motion field. Then, by fixing the motion field and model parameters, they solve a segmentation problem. Two processes are interleaved until convergence. They also observed that HCF provides a better performance than ICM.

In general, a spatial segmentation contains too many regions. The fusion operation implies re-estimation for all adjacent regions. Furthermore based on the experience from previous work at INRS [27, 9], if a pixel is given a wrong label, it's difficult to correct it simply by a contour adjustment process. That's why we are looking for new approach which could take into account the luminance/color, motion and disparity information and give directly a motion/disparity segmentation.

We have seen that MRF model provides a natural way to model segmentation based on a number of sources. In many existing works, interesting models have been established, but in practical implementations interleaved estimation and segmentation are typically used.

### 2.4.5 Minimum description-length

The *minimum description-length* (MDL) criterion [39] demands the shortest description of image data using a chosen language. In information theory, the "length" of a description is measured by the number of bits to represent information.

Leclerc [29] adopted this criterion to formulate a luminance intensity segmentation problem. He showed that in some cases MDL criterion is equivalent to the MAP criterion. In terms of the segmentation problem, image information is described by a selected segmentation model (piecewise-constant or piecewise-smooth) and the description of the image data given the segmentation model (errors). An interesting point here is that a label of a segment can also carry information about the luminance within a region rather than being a meaningless tag.

In our project, in the first step, to extract an initial disparity segmentation we model a disparity region as a piecewise-constant region (Section 3.1).

In general, the prior probability of the segmentation is unknown. When we define the potential function, we always follow the MDL criterion, e.g., a segmentation with the simplest contour is preferred.

## 2.5 Optimization algorithms

To solve an optimization problem like (1) is difficult; standard gradient-based or descent algorithms cannot provide a global optimum. We review some heuristic optimization algorithms here. Both simulated annealing (SA), and highest confident first (HCF) employ Gibbs distribution to model the local probability.

Simulated annealing was first reported in [32] and then in [25]. Geman and Geman [12] developed an annealing algorithm for image restoration. They proved mathematically that the minimum can be attained if the temperature $T(t) \rightarrow 0$ as $t \rightarrow \infty$ and if the annealing process satisfies an "annealing schedule". Unfortunately, it takes too long to satisfy the schedule. Iterated conditional modes (ICM), which is a deterministic procedure can be considered as a special case of SA when the temperature is set to zero.

Chou and Brown applied their Highest Confidence First (HCF) [8] technique to a segmentation problem [7, 6]. Essentially HCF is a deterministic gradient descent method. The, *augmented label set* contains, besides all the labels for which the labeling problem is defined $L = \{l_1, l_2, ..., l_k\}$, an extra "uncommitted" state $l_0$; $\bar{L} = L \bigcup \{l_0\}$. A site $\mathbf{x}$ is committed to a label $l_i$ at step $t$ if $Z(\mathbf{x}) = l_i$, and it is uncommitted if $Z(\mathbf{x}) = l_o$. Once a site has been committed to a label, it cannot nullify its commitment, but it is allowed to change its commitment to other labels in $L$. In HCF, every site starts in the uncommitted state. At any instant, only the site with the highest confidence is changing their state, i.e., the least stable ones with respect to the current configuration, are allowed to change its states.

A continuation method embeds the objective function $\mathcal{U}(Z)$ in a family of functions $\mathcal{U}(\epsilon, Z)$ for which there is a single local minimum at some large $\epsilon$, and for which the local minima converge to that of $\mathcal{U}(Z)$ as $\epsilon$ approaches zero. Leclerc [29] developed a continuation algorithm for luminance segmentation problem. He used the function $e(\epsilon, x) = \exp(-\frac{x^2}{\epsilon^2})$ to approach the Kronecker delta $\delta(x)$ which is typically used in the segmentation model.

## 2.6 Robust Statistics

The field of robust statistics [17, 23] has been developed to address the fact that parametric models of classical statistics are often approximations of the phenomena being modeled. In particular, the field addresses how to handle *outliers*, gross error due to the violations of assumptions about the model.

In [17] Hampel identified the main goals of robust statistics as follows

1. to describe the structure best fitting the bulk of the data,
2. to identify deviating data points (outliers) or deviating substructures for further treatment, if desired.

In a fitting problem, the objective is to find a set of parameters $\mathbf{a}$ for a parametric function $f$, given a set of observation data $O$ so that the residual errors are minimum:

$$\min_{\mathbf{a}} \sum_{o \in O} \mathcal{E}(o - f(o, \mathbf{a})),$$

where $\mathcal{E}$ is a function which measures the *residual* errors. In general, $\mathcal{E}$ is an even function and increasing in $[0, \infty)$. The quadratic function is typically used, and the problem is known as the standard least-squares estimation problem.

An estimation is said to be robust if the solution is relatively insensitive to small deviations for the bulk of the data or to large deviations for a few data points. The least-squares approach is generally not robust because it assigns a high weight to the outliers (a simple example in [3]). To increase robustness, an estimator must be more forgiving about outlying measurements. It means that the derivative of the function $\mathcal{E}(x)$ tends to zero as $x$ tends to $\infty$. There are a number of robust estimators enumerated in [3]. In [3], Black established a robust optical flow estimation framework by replacing the least-squares estimator in the standard formulation by robust estimators.

# 3 Problem

## 3.1 Motivation

We have reviewed existing work related to object-oriented stereoscopic video coding, especially the key problem of motion- and disparity-based segmentation.

There is a consensus that motion/disparity estimation and segmentation must be simultaneously addressed to attain the best accuracy. To apply directly a clustering technique to a pre-estimated motion/disparity is not the best choice because the estimation precess does not take into account the segmentation information. Most approaches begin with a luminance segmentation, and then try to modify it by sequentially applying a series of criteria.

By contrast, MRFs provide an easy way to establish a general framework to deal with the estimation/segmentation problem. Such a model can simultaneously take into account multiple sources, e.g., motion, disparity and luminance/color. Many models have been developed and there is a trend to make models more complete, therefore more complicated. Nevertheless, even for the simplest model, the corresponding optimization problem has not been solved satisfactorily. All optimization methods must start with a good initial state to attain the global optimum. The existing methods typically take a luminance-based segmentation as the initial segmentation, and then try to refine it. In general, this segmentation is refined locally. Most approaches fall back into the cycle “intensity segmentation, then estimation and then segmentation adjustment”.

Our objective is to adopt a MRF model to establish a general framework for segmentation based on a number of sources, such as motion, disparity and luminance/color. We first establish our approach using MRF models (formulation (1)). We define potential functions that reflect relationship between sources and we use continuation method to directly extract an initial motion and/or disparity-based segmentation. Then we introduce robust estimators into our formulations.

We begin with disparity which is the simplest case if a parallel camera setup is assumed.

## 3.2 Disparity estimation and segmentation

In the case of joint estimation and segmentation of disparity, a maximum *a posterior* probability estimator will maximize the probability $p(\mathbf{d}, s|I_l^t, I_r^t)$ where $\mathbf{d}$ is the field of disparities and $s$ is the segmentation field. The random field $s$ is generally modeled as a random field taking values from an integer set $\{1,,2\ldots,k\}$. If a parallel camera setup is assumed, only the horizontal disparity $d$ will be taken into account. Similarly to (1), it is equivalent to the problem

$$\{s^*, d^*\} = \arg\max_{\{s,d\}} p(I_r|I_l, d, s)p(d|I_l, s)p(s|I_l).$$

The superscripts are omitted to simplify the expression.

The equivalent energy optimization is

$$\{s^*, d^*\} = \arg\min_{\{s,d\}} \mathcal{U}(I_r|I_l, d, s) + \mathcal{U}(d|I_l, s) + \mathcal{U}(s|I_l). \tag{10}$$

In our development, we are interested in the so called *redescending* robust estimators which employ even functions $\mathcal{E}(x)$ to measure the residual errors (Section 2.6). The function $\mathcal{E}(x)$ is strictly increasing for $x > 0$, $f(0) = 0$ and $f'(x) \to 0$ as $x \to \infty$. We will use $\mathcal{E}$ to denote such a function. Then we have

$$\mathcal{U}_o(d) \triangleq \mathcal{U}(I_r|I_l, d, s) = \sum_{(X,Y)} \mathcal{E}_0(I_r(X + d, Y) - I_l(X, Y)) \tag{11}$$

In most approaches using MRF models, the least-squares estimator is used instead.

From the estimators enumerated in [3], we have chosen the Lorentzian function for $\mathcal{E}_0$

$$\mathcal{E}(x) = \log(1 + \frac{x^2}{2\theta^2}),$$

where $\theta$ is a scale parameter, because it is continuously differentiable. This property is extremely important since we adopt the continuation method to solve our optimization problem.

The second term in (10) gives a smoothness constraint on the disparity field within a disparity region. In the present project, we consider only 2-element cliques $c = \{(X, Y), (X', Y')\}$. If we define

$$\delta(x) = \begin{cases} 1 \text{ if } x = 0, \\ 0 \text{ otherwise} \end{cases},$$

then

$$\begin{aligned} \mathcal{U}_s(d, s) &\triangleq \mathcal{U}(d|I_l, s) = \sum_{c\in\mathcal{C}} \mathcal{V}_c(d|I_l, s), \\ \mathcal{V}_c(d|I_l, s) &= \alpha\mathcal{E}_1(d(X, Y) - d(X', Y'))\delta(s(X, Y) - s(X', Y')). \end{aligned}$$

The third term in (10) describes *a priori* knowledge about the segmentation:

$$\begin{aligned} \mathcal{U}_a(s) &\triangleq \mathcal{U}(s|I_l) = \sum_{c\in\mathcal{C}} \mathcal{V}_c(s|I_l) \\ \mathcal{V}_c(s|I_l) &= (\beta\mathcal{G}(I_l(X, Y) - I_l(X', Y')) + \gamma)(1 - \delta(s(X, Y) - s(X', Y'))). \end{aligned} \tag{12}$$

The term $\mathcal{G}(x)$ is a function that gives a penalty to neighbors having similar intensity values in order to increase the weight associated with the prior energy. Mathematically speaking, $\mathcal{G}(x)$ is an even function and increasing in $(-\infty, 0]$. The goal of $\mathcal{G}(x)$ is to increase the penalty in constant-intensity areas if a segmentation boundary is introduced. There is a similar potential definition in [1]. The term $\gamma(1 - \delta(s(X,Y) - s(X',Y')))$ reflecting the complexity of a region, is used in almost all approaches treating jointly motion estimation and segmentation problems. Although the optimum segmentation of an image based on rate-distortion criterion does not necessarily coincide with the real contours of objects [45], we think that a segmentation approximating the real contours should be more stable temporally. Note that coding of shape takes a large portion of bits in an object-based coder.

Finally, the optimization problem can be described as follows:

$$\min_{\{s,d\}} \mathcal{U}_o(d) + \mathcal{U}_s(s,d) + \mathcal{U}_a(s). \tag{13}$$

It's not easy to solve the above optimization problem; almost all optimization methods need an initial state that is near enough to the optimum. As we mentioned before, the existing approaches either use a first guess of disparity and segmentation or sequentially follow the steps: spatial segmentation, then estimation, and then segment adjustment. In our development, we want to use neither a spatial segmentation as initial segmentation nor a segmentation obtained from an initial disparity field computed directly from intensities.

It's clear that if we have a good disparity-based segmentation the whole estimation procedure would be simpler. The random field $s$ is generally modeled as a random field taking values from an integer set $S = \{1, , 2 \ldots, k\}$. Here we prefer a set with finite number of members, not necessarily an integer set. It changes nothing because the values are just tags. With any two finite sets of real numbers with the same number of members (13) defines the same optimization problem. We choose one such set that gives an approximation of disparity field. In the extreme case, if the images contain piecewise-constant disparity surfaces, it could happen exactly that $s = d$. Although not all surfaces have constant depth, this assumption is still reasonable for images that don't contain many surfaces with huge depth range.

We can see that

$$\begin{aligned} s^* \quad &= \quad \arg\min_{\{s,d\}} \mathcal{U}_o(d) + \mathcal{U}_s(s,d) + \mathcal{U}_a(s) \\ \text{subject to} \quad & \quad s = d \\ &= \quad \arg\min_{s} \mathcal{U}_o(s) + \mathcal{U}_a(s). \end{aligned} \tag{14}$$

The solution to problem (14) will give us a segmentation directly from the disparity information. Since $s$ is real, the only discrete term in the problem is the $\delta$ function. If the continuation technique developed in [29] is employed, a gradient-based algorithm can be applied to compute $s$.

We use the same function to approach $\delta$ as in [29]. Let $e(\epsilon, x) = e^{-\frac{x^2}{\epsilon^2}}$. Then,

$$\lim_{\epsilon \to 0} e(\epsilon, x) = \delta(x).$$

The approach of continuation is to replace $\delta$ by $e(\epsilon, x)$ and then let $\epsilon$ go to 0. For every $\epsilon$, we have an optimization problem

$$\begin{aligned} \min_{\{s \overset{\triangle}{=} s(\epsilon)\}} \quad & \sum_{(X,Y)} \mathcal{E}_0(I_r(X+s,Y) - I_l(X,Y)) \\ + \quad & \sum_{c \in \mathcal{C}} (\beta \mathcal{G}(I_l(X,Y) - I_l(X',Y')) + \gamma)(1 - e(\epsilon, s(X,Y) - s(X',Y'))). (15) \end{aligned}$$

The term $e(\epsilon, x)$ is usually referred to as the Leclerc estimator because for any given $\epsilon$, $e(\epsilon, s(X,Y) - s(X',Y'))$ plays the role of an estimator.

The algorithm runs as follows:

1. Choose an initial $\epsilon_0$, a positive constant $\lambda < 1$ and a threshold $\epsilon^*$.
2. If $\epsilon_i > \epsilon^*$, solve the problem (15) and update $\epsilon_i : \epsilon_{i+1} = \epsilon_i \times \lambda$. Otherwise stop.

To avoid being trapped in a local optimum, multiresolution optimization should be adopted. Multiresolution techniques calculate the estimate at various levels of spatial resolution. From one level to another, images are filtered by a Gaussian filter and then sub-sampled horizontally and vertically by 2:1. The situation is depicted in Figure 4.

Finally we return to the problem (13). $s^*(\epsilon^*)$ is not only a good state for $s$ but also for $d$. In theory, as $\epsilon \to 0$, the limit of $s^*(\epsilon)$ tends to a segmentation. For practical proposes, it's only a quasi-segmentation; the neighboring values are similar but, in general, not identical. $s$ has to be quantized, and then a label is assigned to $s$. A uniform quantizer is not suitable because we have observed during experiments that smaller values of $s$ (around 1) are much more reliable than greater values (e.g., around 5).

## 3.3 Motion estimation and segmentation

The only difference between motion and disparity/segmentation is that a motion vector has two components $\mathbf{v} = (v_x, v_y)$. We can easily formulate the optimization problem

$$\min_{\{\mathbf{v},s\}} \quad \sum_{(X,Y)\in\Lambda} \mathcal{E}_0(I^{t+1}(X+v_x, Y+v_y) - I^t(X,Y)) \tag{16}$$

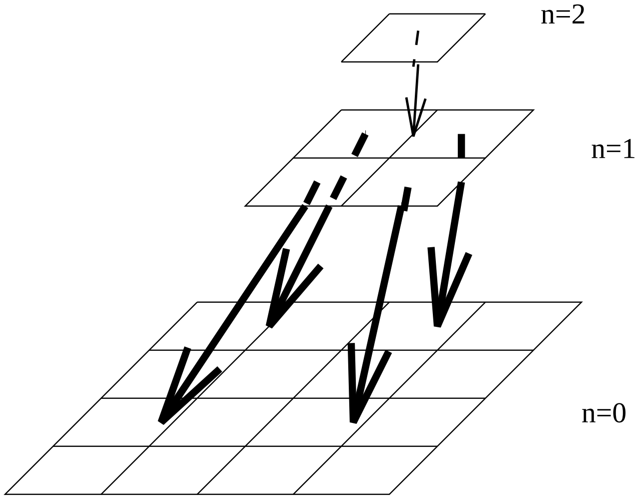

Figure 4: Multi-resolution structure

$$
\begin{aligned}
&+ \quad \alpha \sum_{c \in \mathcal{C}} \mathcal{E}_1(||\mathbf{v}(X,Y) - \mathbf{v}(X',Y')||)\delta(s(X,Y) - s(X',Y')) \\
&+ \quad \sum_{c \in \mathcal{C}} (\beta \mathcal{G}(I^t(X,Y) - I^t(X',Y')) + \gamma)(1 - \delta(s(X,Y) - s(X',Y'))).
\end{aligned}
$$

Solving directly this problem is also difficult. Following the same idea, we look for a good initial segmentation which can approximate a piecewise-constant motion field. The segmentation is obtained based on motion vectors, more precisely on their two components $v_x$ and $v_y$. The final segmentation labels will be computed from two components $s_x$ and $s_y$. Because $s_x$ and $s_y$ are different, they don't have the same contribution to the final segmentation. For simplicity however, we assume that they contribute equally to the final segmentation, and confine the problem to the quantization and optimization steps. This means $s$ can be obtained by superposing $s_x$ onto $s_y$. Formally,

$$s(X,Y) = s(X',Y') \Longleftrightarrow s_x(X,Y) = s_x(X',Y') \text{ and } s_y(X,Y) = s_y(X',Y').$$

To obtain a motion-based segmentation we have to solve the optimization problem

$$
\begin{aligned}
\min_{\{s_x,s_y\}} \quad & \sum_{(X,Y)\in\Lambda} \mathcal{E}_0(I^{t+1}(X+s_x, Y+s_y) - I^t(X,Y)) \\
+ \quad & \sum_{c \in \mathcal{C}} (\beta \mathcal{G}(I^t(X,Y) - I^t(X',Y')) + \gamma) \\
& \times (2 - \delta(s_x(X,Y) - s_x(X',Y')) - \delta(s_y(X,Y) - s_y(X',Y'))).
\end{aligned} \tag{17}
$$

Once estimated, the vectors $(s_x, s_y)$ must be quantized to give an initial segmentation for $s$.

## 3.4 Optimization problem

After solving problems (14) and (17), we have a good initial segmentation as well as motion and disparity fields for problems (16) and (13). Take $s$ $(s_x, s_y)$ as the initial $d$ $(v_x, v_y)$. Because we can never really let $\epsilon \to 0$, a quantization process should be performed to assign labels to $s$. A scalar (vector) quantizer will be developed to obtain discrete labels. The quantized values will give a good approximation to the segmentation sought.

To solve problems (13) and (16), I anticipate to develop an algorithm similar to HCF optimization [6]. The reason that I want to develop an HCF-style algorithm is that this technique provides a way to give a confidence measure to every change. I intend to develop a *local stability measure* (confidence measure) different from that described in [6]. Although I have not formulated the local stability measure yet, the measure should, based on our experiments, depend on

- the amplitude of motion (disparity) vectors,
- the distance of a pixel from the boundary of region it belongs to.

As for the continuation of $d$, I need to develop it further, i.e., discrete or continue state space.

## 3.5 Region-based simultaneous motion/disparity estimation

In this section we establish our complete model for joint (disparity, motion) estimation and segmentation. Our sources are a stereopair $I_l^t, I_r^t$ and a pair of consecutive frames (fields) $I_l^t, I_l^{t+1}$.

For this problem, $\mathbf{z} = \{\mathbf{v}_l^t, d^t, s_l^t, s_d^t, \}$ where $s_d^t$ stands for disparity-based segmentation at $t$ while $s_l^t$ stands for the motion segmentation at instant $t$. We must maximize

$$\max_{\mathbf{z}} P(\mathbf{z}|I_l^t, I_l^{t+1}, I_r^t). \tag{18}$$

Factoring the probability, we have

$$\begin{aligned}
P(\mathbf{z}|I_l^t, I_l^{t+1}, I_r^t) &= \frac{P(\mathbf{z}, I_l^t, I_l^{t+1}, I_r^t)}{P(I_l^t, I_l^{t+1}, I_r^t)} \\
&= \frac{P(I_l^{t+1}|s_d^t, s_l^t, \mathbf{v}_l^t, d^t, I_r^t, I_l^t)}{P(I_l^t, I_l^{t+1}, I_r^t)} \\
&\times \; P(I_r^t|d^t, \mathbf{v}_l^t, s_l^t, s_d^t, I_l^t) \\
&\times \; P(\mathbf{v}_l^t, d^t|s_l^t, s_d^t, I_l^t) \\
&\times \; P(s_l^t|s_d^t, I_l^t) \times P(s_d^t|I_l^t).
\end{aligned}$$

The equivalent problem in terms of energies is

$$\min \quad \mathcal{U}(I_l^{t+1}|s_d^t, s_l^t, \mathbf{v}_l^t, d^t, I_r^t, I_l^t)$$

$$
\begin{aligned}
&+ \quad \mathcal{U}(I_r^t | d^t, \mathbf{v}_l^t, s_l^t, s_d^t, I_l^t) \\
&+ \quad \mathcal{U}(\mathbf{v}_l, d^t | s_l^t, s_d^t, I_l^t) \\
&+ \quad \mathcal{U}(s_l^t | s_d^t, I_l^t) + \mathcal{U}(s_d^t | I_l^t).
\end{aligned}
$$

The corresponding potentials are defined below.

### 3.5.1 Observation models

Observation models are established to reflect the quality of the compensation of luminance values by motion and disparity fields:

$$
\begin{aligned}
\mathcal{U}_o(\mathbf{v}_l^t, d^t) &\triangleq \mathcal{U}(I_l^{t+1} | s_d^t, s_l^t, \mathbf{v}_l^t, d^t, I_r^t, I_l^t) + \mathcal{U}(I_r^t | d^t, \mathbf{v}_l^t, s_l^t, s_d^t, I_l^t) \\
&= \sum \mathcal{E}_0(I_l^{t+1}(X + v_{x,l}, Y + v_{y,l}) - I_l^t(X, Y)) \\
&+ \sum \mathcal{E}_0(I_r^t(X + d^t, Y) - I_l^t(X, Y)).
\end{aligned}
$$

### 3.5.2 Smoothness constraints

Smoothness constraints reflect the *a priori* knowledge about disparity and motion fields within every region:

$$
\begin{aligned}
\mathcal{U}_s(\mathbf{v}_l^t, d^t, s_l^t, s_d^t) &\triangleq \mathcal{U}(\mathbf{v}_l^t, d^t | s_l^t, s_d^t, I_l^t) \\
&= \sum_{c \in \mathcal{C}} \alpha \mathcal{E}_1(||\mathbf{v}_l^t(X, Y) - \mathbf{v}_l^t(X', Y')||) \delta(s_l^t(X, Y) - s_l^t(X', Y')) \\
&+ \sum \alpha \mathcal{E}_1'(d^t(X, Y) - d^t(X', Y')) \delta(s_d^t(X, Y) - s_d^t(X', Y'))
\end{aligned}
$$

Because of different characteristics of disparity and motion fields we have to choose different estimators for them. In case of the Lorentzian estimator

$$
\mathcal{E}(x) = \log(1 + \frac{x^2}{2\theta^2}),
$$

we will choose different $\theta$.

### 3.5.3 Segmentation constraint

Here we have to define a function reflecting the relationship between $s_d^t$ and $s_l^t$. Our assumption is that any disparity region is contained in a motion region. In other words, a motion boundary must coincide with a disparity (depth) boundary but not *vice versa.* For a 2-element clique, we define:

$$
\mathcal{V}_c^c = \begin{cases} 1 & \text{if } s_d^t(X, Y) = s_d^t(X', Y'), s_l^t(X, Y) \neq s_l^t(X', Y') \\ 0 & \text{otherwise} \end{cases} .
$$

Then, we formulate an energy term using the $\delta$ function

$$\begin{aligned}\mathcal{U}_c(s_d^t, s_l^t) &\stackrel{\triangle}{=} \mathcal{U}(s_l^t | s_d^t, I_l^t) \\ &= \sum_{c \in \mathcal{C}} \zeta \mathcal{V}_c^c \\ &= \sum_{c \in \mathcal{C}} \zeta \delta(s_d^t(X,Y) - s_d^t(X',Y'))(1 - \delta(s_l^t(X,Y) - s_l^t(X',Y')))\end{aligned}$$

This term not only reflects an assumption about the segmentations but also joins the problems of motion estimation/segmentation and of disparity estimation/segmentation.

### 3.5.4 A priori model

This model quantifies the complexity of the disparity, and therefore motion, boundaries (regions):

$$\begin{aligned}\mathcal{U}_a(s_d^t) &\stackrel{\triangle}{=} \mathcal{U}(s_d^t | I_l^t) \\ &= \sum_{c \in \mathcal{C}} \mathcal{V}_c(s_d^t | I_l) \\ \mathcal{V}_c(s_d^t | I_l) &= (\beta \mathcal{G}(I_l(X,Y) - I_l(X',Y')) + \gamma)(1 - \delta(s_d^t(X,Y) - s_d^t(X',Y'))).\end{aligned}$$

### 3.5.5 Total energy

Finally, we can formulate the minimization of the whole objective function as follows:

$$\min_{\mathbf{z}} \mathcal{U}_o(\mathbf{v}^t, d) + \mathcal{U}_s(\mathbf{v}^t, d^t, s_d^t, s_l^t) + \mathcal{U}_c(s_d^t, s_l^t) + \mathcal{U}_a(s_d^t). \tag{19}$$

We will develop and compare two approaches. The first one will solve sequentially the problems (14), (17) and (19). Formally,

1. solve independently (14) and (17).
2. quantize and fuse $s_x, s_y$ to obtain $s_l^t$ and quantize the initial disparity segmentation $s_d$ to obtain $s_d^t$.
3. using $s_x, s_y, s_d$ as initial $v_{l,x}^t, v_{l,y}^t, d^t$ respectively, and using quantized $s_d^t$ and $s_l^t$ as initial segmentations , apply the developed HCF algorithm to (19).

Above, initial $v_{l,x}^t, v_{l,y}^t, s_l^t$ and $d^t, s_d^t$ to be used in problem (19), were computed independently. We believe that at an increased computational cost they can be computed jointly and therefore provide a more reliable result. In the second approach, we use the continuation method (as in Section 3.1 and 3.2) to compute the initial segmentations. Precisely, we Let $s_d^t = d^t, s_{l,x}^t = v_{l,x}^t$, and $s_{l,y}^t = v_{l,y}^t$ in each energy of 19. (The segmentation $s_l^t$ is determined by the two components $s_{l,x}^t, s_{l,y}^t$ as described in Section 3.3.) For any given $\epsilon$, we have to solve the optimization problem

$$\min_{\{s_{l,x}^t, s_{l,y}^t, s_d^t\}} \mathcal{U}_o(s_{l,x}^t, s_{l,y}^t, s_d^t) + \mathcal{U}_c(s_d^t, s_{l,x}^t, s_{l,y}^t, \epsilon) + \mathcal{U}_a(s_d^t, \epsilon). \tag{20}$$

where

$$\begin{aligned}
&\mathcal{U}_c(s_d^t, s_{l,x}^t, s_{l,y}^t, \epsilon) \\
= \;& \zeta \sum_{c \in \mathcal{C}} \exp\left(\epsilon, (s_d^t(X,Y) - s_d^t(X',Y'))\right) \\
& \times (2 - \exp\left(\epsilon, s_{l,x}^t(X,Y) - s_{l,x}^t(X',Y')\right) - \exp\left(\epsilon, s_{l,y}^t(X,Y) - s_{l,y}^t(X',Y')\right)),
\end{aligned}$$

and

$$\mathcal{U}_a(s_d^t, \epsilon) = \sum_{c \in \mathcal{C}} (\beta \mathcal{G}((I_l(X,Y) - I_l(X',Y')) + \gamma)(1 - \exp(\epsilon, s_d^t(X,Y) - s_d^t(X',Y'))).$$

The second approach can be performed in the following three steps:

1. solve (20) until $\epsilon \leq \epsilon^*$,
2. quantize and fuse $s_{l,x}^t, s_{l,y}^t$ to obtain $s_l^t$ and quantize initial disparity segmentation,
3. using $s_{l,x}^t, s_{l,y}^t, s_d^t$, as initial $v_{l,x}^t, v_{l,y}^t, d^t$ respectively, and using quantized $s_d^t$ and $s_l^t$ as initial segmentations, apply the developed HCF algorithm to the problem (19).

We hope the second approach will give better quality of results because the segmentations are adjusted adaptively during the continuation on $\epsilon$, while in the first approach the segmentation constraint will only function in the final step. The first approach has the advantage of a lower computational complexity.

## 3.6 Performance measures

The ideal way to evaluate the performance of the proposed approach would be to simulate a complete object-based encoder, employ the same coding method to encode a stereo sequence based on the estimation and segmentation results obtained from different approaches and compare bit rates given a measure of image quality. It requires a complete implementation of an object-based encoder. Such an implementation is beyond the scope of this thesis.

In the project, we will evaluate separately estimation and segmentation quality instead. We hope to have better estimation which can be verified by examining the quality of the prediction based on the estimation result.

As far as the quality of segmentation is concerned, it's not easy to establish a criterion to evaluate the quality of segmentation. Only for synthetic sequences, we know exactly the disparity/motion segmentation. Experiments will be conducted on synthetic sequences to verify whether the segmentation properly classifies real disparity/motion objects.

# 4 Initial results

In this section, we present results that we have obtained to date.

## 4.1 Rectification

As we know there are two types of stereo camera setups, toed-in and parallel cameras. Parallel cameras have the following advantages: simple geometry, therefore simple processing, less distortion, perfect perception possible. By contrast, toed-in cameras suffer from more distortion. Their complicated geometric structure makes stereo sequence processing more tedious. Unfortunately, the angle of convergence is inevitable in a practical stereo camera system. A rectification technique permits to adjust the convergence by projecting the images obtained by a toed-in camera onto a virtual parallel stereo camera setup. In Section 2.2, we have mentioned that there is no vertical disparity in a parallel camera system. We have developed a rectification process (details are described in a technical report [57]) to adjust our stereo sequences. Therefore, in our work we always assume that there is no vertical disparity.

## 4.2 Initial results

To date, we have studied the problems (14) and (17). We used a simple descent algorithm to solve the optimization problem (15) for a given $\epsilon$.

First, we took the least-squares estimator for $\mathcal{E}_0$ and then replaced it by a Lorentzian estimator. In the experiments, the penalty function $\mathcal{G}$ was fixed as $e(\theta, x)$.

Several motion, estimation/segmentation experiments have been conducted on fields 0 and 2 of the sequence "autoroute". The experiments were performed for the following cases:

- fixing $\gamma = 0$, try different values for $\beta$,
- fixing $\beta = 0$, try different values for $\gamma$,
- use least-squares estimator,
- use Lorentzian estimator.

A large value of $\gamma$ emphasizes complexity term hence assures simpler contours. A large $\beta$ penalizes different tags in uniform areas hence forces contours to adopt objects' real shape. The experiments show that

- using a robust estimator the results can be improved,
- a good quality quasi-segmentation can be obtained,

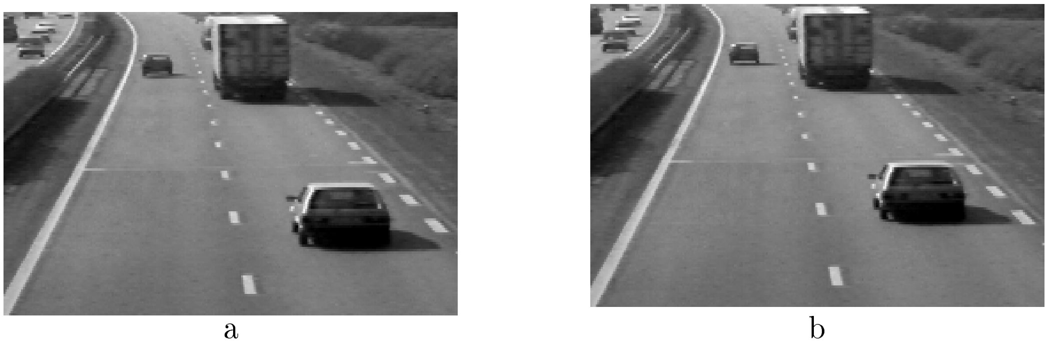

Figure 5: Sequence "autoroute": field 0 and field 2

- a high weight $\beta$ forces the segmentation to adapt to the luminance information,
- multiresolution optimization not only significantly reduces the computation time but also improves the estimation; large vectors (6 pixels) are properly estimated.

In the next step we should try to find a good combination of $\beta$ and $\gamma$.

The original fields number 0 and 2 are shown in Figure 5. The results obtained are reported in Figures 6, 7 and 8. In Figure 6, the results are obtained using the least-squares estimator. Figure 6(a) corresponds to the case of $\beta = 0$ is while Figure 6(b) corresponds to the case $\gamma = 0$.

The results obtained using the Lorentzian estimator are reported in Figure 7. Figure 7(a) corresponds to the case $\beta = 0$ while Figure 7(b) corresponds to the case $\gamma = 0$.

We can see the quasi-segmentations $s_x$ and $s_y$ in Figure 8 represented as gray level values. The white pixels correspond to positive values while the black pixels correspond to negative values. Figure 8(a) is obtained from the horizontal component while Figure 8(b) is obtained from the vertical component.

# References

[1] A. Alatan and L. Onural, "Object-based 3d motion and structure estimation," in *International Conference on Image Processing*, 1995.

[2] M. Anderberg, *Cluster Analysis for Applications.* Academic Press New York, 1973.

[3] M. Black, *Robust Incremental Optical Flow.* PhD thesis, Yale University, 1992.

[4] M. Chang, A. Tekalp, and I. Sezan, "Simultaneous motion estimation and segmentation." IEEE Transactions on Image Processing, in print.

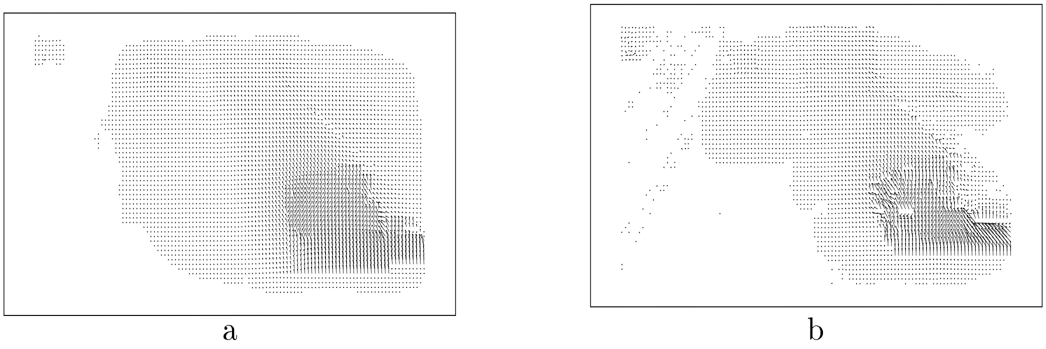

Figure 6: The least-squares estimator: a. $\beta = 0$ b. $\gamma = 0$

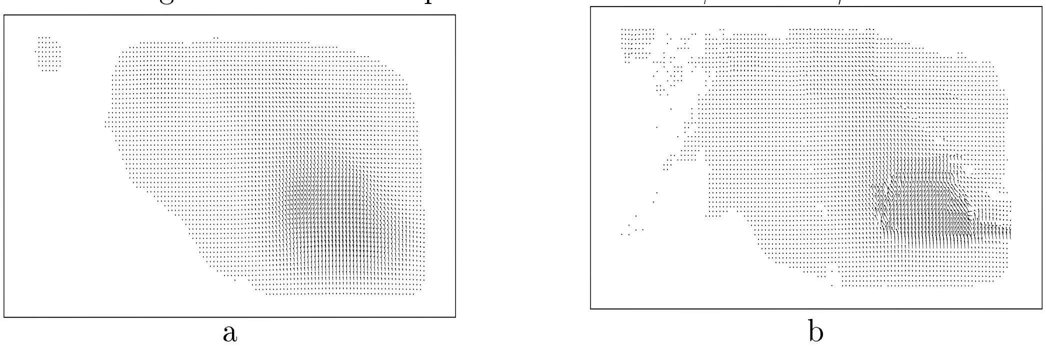

Figure 7: Lorentzian estimator: a. $\beta = 0$ b. $\gamma = 0$

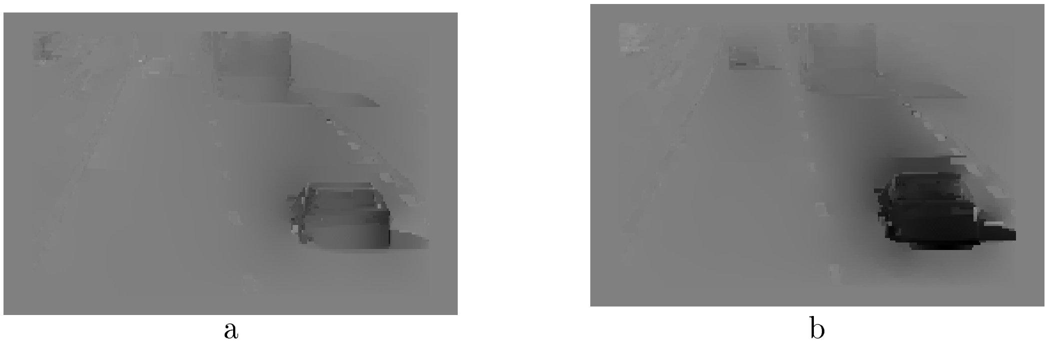

Figure 8: Quasi-segmentation: a. horizontal component b.vertical component

[5] B. Choquet and J. Fournier, "Importance of opto geometrical adjustments for stereoscopic television," in *2nd International Conference on 3D Media Technologies*, 1992.

[6] P. Chou and C. Brown, "The theory and practice of bayesian image labeling," *International Journal of Computer Vision*, vol. 4, pp. 185–210, 1990.

[7] P. Chou and C. Brown, "Multimodal reconstruction and segmentation with markov random fields and hcf optimization," in *Proceedings Image Understanding Workshop, Cambridge, MA*, February 1988.

[8] P. Chou and R. Raman, "On relaxation methods based on markov random fields," tech. rep., Computer Science Dept. Rochester Univ., 1987.

[9] V. Dang, A. Mansouri, and J. Konrad, "Motion estimation for region-based video coding," in *International Conference on Image Processing, Washington*, 1995.

[10] N. Diehl, "Object-oriented motion estimation and segmentation in image sequences," *Signal Processing: Image Communication*, vol. 3, pp. 23–56, 1991.

[11] O. Faugeras, *Three-dimensional Computer Vision: A geometric viewpoint.* The MIT Press, 1993.

[12] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions and the bayesian restoration of images," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. PAMI-6, no. 6, pp. 721–741, 1984.

[13] P. Gerken, "Object-based analysis-synthesis coding of image sequence at very low bit-rates," *IEEE Trans. Circ. Syst. Video Technol.*, vol. 4, June 1994.

[14] M. Gilge, T. Engelhardt, and R. Mehlan, "Coding of arbitrarily shaped image image segments based on a generalized orthogonal transform," *Signal Processing: Image Communication*, vol. 1, pp. 153–180, 1989.

[15] N. Grammalidis, S. Malassiotis, D. Tzovaras, and M. Strinzis, "Stereo image sequence coding based on three-dimensional motion estimation and compensation," *Signal Processing: Image Communication*, vol. 7, 1995.

[16] P. Gunatilake, M. Siegel, and A. Jordan, "Compression of stereo video," in *International workshop on HDTV 1993*, 1993.

[17] F. Hampel, E. Ronchetti, P. Rousseeuw, and W. Stahel, *Robust Statistics: The Approaches Based on Influence Functions.* John Wiley and Sons, New York, NY, 1986.

[18] G. Heckmer, "Redundancy reducing coding of moving object shapes," *Signal processing: Image Communication*, vol. 9, pp. 91–98, 1997.

[19] L. Hodges, “Basic principles of stereographic software development,” vol. 1457, SPIE, 1991.

[20] H. Horn and B.G. Schunck, “Determining optical flow,” *Proc. Artificial Intelligence*, vol. 17, 1981.

[21] M. Hotter, “Object-oriented analysis-synthesis coding based on moving two-dimensional objects,” *Signal Processing: Image Communication*, vol. 2, pp. 409–428, 1990.

[22] M. Hotter and R. Thoma, “Image segmentation based on object oriented mapping parameter estimation,” *Signal Processing: Image Communication*, vol. 15, pp. 351–334, 1988.

[23] P. Huber, *Robust Statistics.* John Wiley and Sons, New York, NY, 1981.

[24] ISO/IEC JTC1/SC29/WG11 N998, *MPEG-4 Proposal package description (PPD)-Revision3(Tokyo Revision)*, July 1995.

[25] S. Kirkpatrick, C. Gelatt, and M. Vecchi, “Optimization by simulated annealing,” *Science*, vol. 220, May 1983.

[26] R. Koch, “Model-based 3D scene analysis from stereoscopic image sequences,” European Workshop on 3D-TV, (Rennes), November 1992.

[27] J. Konrad and V. Dang, “Coding-oriented video segmentation inspired by MRF models,” in *SPIE*, 1996.

[28] A. Kopemik and D. Pele, “Improved disparity estimation for the coding of stereoscopic television,” in *SPIE*, vol. 1818, 1992.

[29] Y. Leclerc, “Constructing simple stable descriptions for image partitioning,” *International Journal of Computer Vision*, vol. 3, 1989.

[30] W. Lee and C. Chan, “Two-dimensional split and merge algorithm for image coding,” in *SPIE*, vol. 2501, 1995.

[31] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” University of California Press, Berkeley, 1967.

[32] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, “Equation of state calculations by fast computing machines,” *J. Chem. Phys*, vol. 21, no. 6, pp. 1087–1092, 1953.

[33] H. Musmann, “Object-oriented analysis-synthesis coding based on source models of moving 2D and 3D-object,” in *IEEE International Conference on Acoustics, Speech, Signal Processing*, vol. I, 1993.

[34] H. Musmann, M. Hotter, and Jorn, “Object-oriented analysis-synthesis coding of moving images,” *Signal Processing: Image Communication*, vol. 1, pp. 117–138, 1989.

[35] J. Ostermann, “Object-oriented analysis-synthesis coding based on the source model of moving rigid 3d objects,” *Signal processing: Image Communication*, vol. 6, no. 5, pp. 143–161, 1994.

[36] J. Ostermann, “Object-oriented analysis-synthesis coding based on the source model of moving flexible 3d objects,” *IEEE Transaction on Image Precessing*, vol. 3, pp. 705–711, September 1994.

[37] M. Perkins, “Data compression of stereopairs,” *IEEE Transactions on Communications*, vol. 40, pp. 684–695, April 1992.

[38] A. Puri, R. Kollarits, and B. Haskell, “Stereoscopic video compression using temporal scalability,” in *SPIE*, vol. 2501, 1995.

[39] J. Rissanen, “Minimum-description-length principle,” in *Encyclopedia of Statistics Sciences*, vol. 5, Wiley: New York, 1987.

[40] H. Schiller and M. Hotter, “Investigations on color coding in an object-oriented analysis-synthesis coder,” *Signal Processing: Image Communication*, vol. 5, pp. 319–326, 1993.

[41] C. Stiller and J. Konrad, “On models, criteria and search strategies for motion estimation in image sequences.” submitted.

[42] C. Stiller and B. Hurtgen, “Combined displacement estimation and segmentation in image sequence,” in *Fiber Optic network and Video Compression, Berlin, Germany*, 1993.

[43] C. Stiller, “Object-oriented video coding employing dense motion fields,” in *IEEE International Conference on Acoustics Speech Signal Processing 94*, 1994.

[44] C. Stiller, “Object-based estimation of dense motion field,” *IEEE Transactions on Image processing*, vol. 6, February 1997.

[45] K. Stuhlmuller, A. Salai, and B. Girod, “Rate-constrained contour-representation for region-based motion compensation,” in *VCIP-96*, 1996.

[46] A. Tamtaoui and C. Labit, “Constrained disparity and motion estimations for 3dtv image sequence coding,” *Signal Processing: Image Communication*, no. 4, 1991.

[47] A. Tekalp, *Digital Video Processing.* Prentice Hall, 1995.

[48] R. Tsai and T. Huang, "Estimating three-dimensional motion parameters of a rigid planar patch," *IEEE Transaction on Acoustics, Speech, and Signal Processing*, vol. ASSP-29, December 1981.

[49] B. Tseng and D. Anastassiou, "Compatible video coding of stereoscopic sequences using mpeg-2's scalability and interlaced structure," in *International Workshop on HDTV*, October 1994.

[50] B. Tseng and D. Anastassiou, *A multi-viewpoint digital video codec utilizing computer graphics tools and image warping tech. based on perspective constructions of intermediate viewpoint images*. ISO-IEC/JTC1/SC29/WG11, July 1995.

[51] B. Tseng and D. Anastassiou, "A theoretical study on an accurate reconstruction of multiview images based on the Viterbi algorithm," in *International Conference on Image Processing*, 0-8186-7310-9/95, 1995.

[52] D. Tzovaras, N. Grammalidis, and M. Strintzis, "Joint three-dimensional motion/disparity segmentation for object-based stereo image sequence coding," *Optical Engineering*, vol. 35, pp. 137–144, January 1996.

[53] D. Walker and K. Rao, "Improved pel recursive motion compensation," *Proc. IEEE Trans. Comm.*, vol. COM-32, 1984.

[54] W. Woo and A. Ortega, "Stereo image compression with disparity compensation using MRF model," in *VCIP*, 1996.

[55] W. Woo and A. Ortega, "Stereo image compression based on disparity field segmentation," in *SPIE*, vol. 3024, 1997.

[56] A. Woods, T. Docherty, and R. Koc, "Image distortions in stereoscopic video system," in *Stereoscopic Display and Applications IV*, 1993.

[57] C. Yang, "Geometric models in stereoscopic video," Tech. Rep. 12, INRS-Telecommunications, 1995.