Object-based Video Segmentation with Region Labeling

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof.ir. K.F. Wakker,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen

op maandag 26 november 2001 om 10.30 uur

door

Ioannis PATRAS

M.Sc. in Informatica
geboren te Thessaloniki, Griekenland

Dit proefschrift is goedgekeurd door de promotor:
Prof.dr.ir. R.L. Lagendijk

Toegevoegd promotor:
Dr. E.A. Hendriks

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof.dr.ir. R.L. Lagendijk,	Technische Universiteit Delft, promotor
Dr. E.A. Hendriks,	Technische Universiteit Delft, toegevoegd promotor
Prof.dr.ir. J. Biemond,	Technische Universiteit Delft
Prof.dr.ir. A. Smeulders,	Universiteit van Amsterdam
Prof.dr.ir. L.J van Vliet,	Technische Universiteit Delft
Dr.drs. L.J.M. Rothkrantz,	Technische Universiteit Delft
Dr. G. Tziritas,	University of Crete

ISBN

This manuscript was edited with Emacs under the RedHat Linux operating system. It is typeset in 10 points Times-Roman by LaTeX. The figures were created using *xfig* and the graphs using *gnuplot*. The bibliography was compiled by BibTeX.

Copyright ©2001 by Ioannis Patras I.Patras@ieee.org
All rights reserved. No part of this publication may be reproduced in any form or by any means, electronic or mechanical, including photocopying, recording, or information storage and retrieval, without written permission of the author.

To my grandparents

Object-based Video Segmentation with Region Labeling

Ioannis Patras

Abstract

In this dissertation we propose three methods for object-based segmentation of image sequences. In all three of them, a fine initial intensity or color segmentation provides a set of segments which are subsequently labeled according to statistical models of the objects' properties. We illustrate the advantages of such an approach in terms of robustness, localization accuracy and computational complexity. The first two methods assume that the kinematic behavior of the objects in the scene can be described with parametric motion models. The first method proposes a two-stage approach: in the first stage the parameters of the motion models are estimated from an independently estimated motion field with a clustering scheme that incorporates motion-specific confidence measures and techniques inspired from robust statistics. In the second stage, the intensity segments are labeled according to a modeling of the distribution of the motion-compensated intensity differences under the extracted motion models and the Markov Random Field modeling of the label field. Cliques are defined between intensity segments and a three-frame approach is adopted in order to deal with occlusions. The second method expresses the spatial and temporal constraints of the labeling problem in a single framework and estimates jointly the label field and the parameters of the motion models by maximizing the *a posteriori* probability of the label field. Spatial and temporal constraints on the label field are expressed in the Markov Random Field framework and a three-frame approach is adopted in order to deal with occlusions. We show that a number of pixel-based methods can be expressed as special cases of the second method and how the latter can be extended to incorporate soft labeling decisions. The third method proposes a semi-automatic approach in which the labeling is based on the modeling of the local color and motion statistical properties of the objects. A user-assisted color segmentation provides an initialization of the label field for the first frame of the sequence, while for the rest of the sequence an initialization of the parameters of the models is provided by a motion-based projection operation in the previous frame. The labeling criterion is the maximization of the joint probability of the label field and the observed color and motion properties and is performed in terms of the statistical representations of the properties of the color segments.

Acknowledgments

There are a lot of people who, in one way or the other, have influenced this thesis. Starting from the begining I would like to thank George Tziritas who, during my M.Sc., guided me in my first scientific steps and established my first contacts with Delft University of Technology.

From the Information and Commulication theory group, I would primarily like to thank my supervisors Emile Hendriks and Inald Lagendijk for their patience, their valuable advises and their moral support at the points that I needed it. I have always enjoyed our conversations; they gave me insight to my work and prepared the ground for new ideas. Their comments were always at the point and even if at some moments, back then, made me mumble, looking back, I am happy they were made. For all the things that I have learned from them in the four years that we worked together I would like to thank them.

I would like to thank also the group of Information and Communication Theory for the warm atmosphere with which it has surrounded me all these years. In particular I would like to thank Cor for our discussions and for his friendship, Eric for saving me more than once with technical advises and of course (maybe the soul of the group) Annett and Ben.

My thanks to Stamatis Vassiliadis should probably be a section by themselves. For the kindness with which you embraced me, for being an invaluable pool of advise and for the pleasure of your company (even in your ’father mood’) thank you.

Before I had completely finished my thesis I started my work at the University of Amsterdam. For their trust in me and for their understanding and support when trying to do two things at the same time I would like to thank Arnold Smeulders and Marcel Worring.

Life of course is not only science and happily I had a lot of friends around me to remind me so. Among them, a thanks to Pyrrhos, Sotiris and Kostas without whose company I am not so sure how I would overcome the shock of the first Dutch winter. A special thank you to Maja, for the pleasure of being a part of my life.

Finally, I would like to thank my family for their love and support with which they surrounded me all these years. And to my grandparents who will always be present, this thesis is dedicated.

Contents

Abstract **I**

Acknowledgments **III**

1 Introduction **1**
1.1 Statement of the problem 2
1.2 Major contributions and organization of the dissertation 4

2 Related Work **9**
2.1 Overview of the Proposed Methods 9
2.2 Indirect Methods for Motion-Based Segmentation 11
2.3 Direct Methods for Motion-Based Segmentation 14
2.4 Tracking of non-rigid objects 19

3 Sequential Motion Estimation and Segmentation **23**
3.1 Intensity segmentation 24
3.2 Motion Hypothesis Extraction 26
  3.2.1 Confidence measures 29
  3.2.2 Clustering 37
3.3 Labeling Using Markov Random Field Theory 45
3.4 Conclusions 50

4 Joint Motion Estimation and Segmentation **53**
4.1 Problem Modeling 55
4.2 MAP estimation 60
  4.2.1 Labeling phase 62
  4.2.2 Motion Estimation phase 62
4.3 Experimental Results 65
  4.3.1 “coast guard” sequence 65
  4.3.2 “train” sequence 70
  4.3.3 “pig” sequence 72
4.4 Conclusions 73

5 Semi-automatic Statistical Segmentation 75
5.1 Feature Extraction and User Interaction . . . 76
5.2 Local Mixture Modeling . . . 79
5.3 Maximization of Joint Probability . . . 81
5.4 Object Projection . . . 84
5.5 Experimental Results . . . 85
5.6 Conclusions . . . 88

6 Discussion 95
6.1 Intensity segments as primary elements . . . 96
6.2 Fully automatic methods . . . 98
6.3 Medium-level user interaction . . . 99
6.3.1 Two-stage motion hypotheses extraction and motion-based labeling . . . 101
6.3.2 Joint motion estimation and segmentation . . . 102
6.3.3 Open issues in motion-based segmentation . . . 104
6.4 Higher level user interaction . . . 104

A Proof of Lemma 1 107

B $|G_s| = 1$ **109**

C Relation with the EM algorithm 111

D Synthetic Image Sequences 113
D.1 “C1” image sequence . . . 113
D.2 “R1” image sequence . . . 114
D.3 “Y1” image sequence . . . 114
D.4 “S5” image sequence . . . 114

E Terminology 117

Bibliography 119

Samenvatting 129

Curriculum vitae 131

Chapter 1

Introduction

Boosted by technological advances in the area of communications and computer engineering, there has been an explosion in the amount of the distributed visual information in the last decade. Digital video and the expected fusion of television and internet services only intensify this trend. This fact and the applications that are continuously emerging advocate the need for the development of a wide range of methods for distributing and processing the available visual information.

Object-based segmentation of image sequences, the area in which this thesis belongs, is one of the issues that often arise in the world of video processing and communications. The goal is the decomposition of a scene in the objects that constitute it. Such a decomposition is provided in terms of the partitioning of each frame of the sequence into a number of disjoint regions each of which is assigned a label which denotes which object is depicted in it[1]. Such a labeling represents each object as a distinct spatio-temporal entity within the image sequence.

By partitioning each frame into segments that correspond to meaningful objects, an object-based representation of a scene can be obtained. Once a partition is available, visual information can be coded, delivered and viewed in terms of its actual contents. This is a step forward from the classical television paradigm that has dominated the world of visual communications up to now and in which the visual information is represented as a sequence of rectangular frames. In that direction, an effort towards an international standard which supports an object-based representation of audiovisual information was launched by the Moving Picture Experts Group (MPEG) [69]. The applications that can benefit from such a semantically meaningful object-based representation are numerous. Let us mention here two examples:

- In the context of the MPEG-4 [70] [71] supported applications, user interaction with the actual contents of the visual information can be achieved in a better, more meaningful and easy manner. In particular new, content-based functionalities such as are object-dependent coding quality or object manipulation for video editing can now be developed. The former can be of importance for example in

---

[1] Unless explicitly stated otherwise, the term *region* is used loosely throughout the thesis to denote a set of spatially connected pixels

video conferencing applications, where a reduction in the required bandwidth can be achieved by a higher quality coding only of regions depicting the face and the hands. On the other hand, video editing, which up to now required that the analysis is performed at the user's end, can be now reduced to the built of user interfaces for the manipulation of the already existing object entities. By the time this thesis was written, such systems were already commercially available (e.g. [26]).

- In the context of MPEG-7 [72] [73], applications that revolve around retrieval of multimedia material in very large distributed databases can benefit from such an object-based representation. The reason is that such a representation offers already a semantical organization of the data (although depending on the application domain the contents may vary significantly). By attaching MPEG-7 descriptors to the objects, we essentially describe the scene in terms of the properties of the objects that are depicted in. This brings closer user queries which are formulated in terms of the object properties to the actual description.

Let us note that both MPEG-4 and MPEG-7 standards carefully avoid standardization of the analysis process and the encoder's side. In relation to the object-based representation that MPEG-4 proposes, that means that no standard exists for extracting the objects in the scene. On the one hand this is because by standardizing only the minimum amount that guarantees interoperability, there is room left for competition on the analysis side. On the other hand, the generality of a standard for object-based segmentation would inevitably be seriously limited. The reason behind that is that there is ambiguity at the semantical level in the definition of the problem. To be more specific, what the result of an object-based segmentation scheme should be, is a question that can be answered in many ways. In the next section, we will briefly discuss the issues related to the this ambiguity. The types of answers that we give will position the methods that were developed in this dissertation in the field of object-based segmentation of image sequences.

1.1 Statement of the problem

In order to formulate methods which decompose a video sequence into the objects that are depicted in it, formal definitions of what constitutes an object are necessary. For humans, this task is directly connected with the cognitive processes involved in the action of seeing. As Parmenides noted, *"What is, is identical to the thoughts of the one who recognizes what it is"*[2] [84]. In this line of thinking, a semantical decomposition of a scene requires human-like reasoning about its contents. However, human-like reasoning about the contents of the scene and its incorporation in an artificial vision system is far from realization.

Current approaches to circumvent the problem attempt formulations that are based on some grouping based on properties extracted from the visual data. The goal is to bridge the gap between the semantical interpretation of the scene by a human user and

---

[2] *το γαρ αυτο νοειν εστιν τε και ειναι*

low level features such as color and motion that can be automatically extracted from the image sequence. Two general directions exist, depending on the amount of *a priori* information that is available about the contents of the scene. In the first, information about the domain is *a priori* available, for example faces [86], human body and hand motion [39] or some medical applications. In such cases, domain-specific models can be employed and the labeling formulated as data fitting into the model in question. As an alternative, more general-purpose models can be employed, whose parameters are trained on data from the domain in question [8]. In both cases, the utilization of the *a priori* knowledge simplifies the problem significantly. On the other hand, such approaches are inherently restricted to the domain for which the models of the scene were specified and trained.

This thesis falls in the second category of the approaches, which attempt a more general treatment of the problem. In the approaches of this category the segmentation is based on homogeneity assumptions on low-level features such as color, texture and motion. Models of the spatiotemporal localization of the objects, of the distribution of their properties and/or of the temporal evolution of these models are commonly employed as useful constraints. The fundamental difference between these models and the domain-specific models employed by the approaches of the first category is the generality of the constraints that are imposed on the objects' properties.

In the methods of this second direction motion proved to be one of the most successful properties for segmentation purposes. The reason is that motion can carry information about the shape, the depth and the physical connectivity of the objects in the scene. These are directly connected to the physical properties of the objects which are depicted in the image sequence and especially the last two can provide strong evidence for the discrimination in different objects. Let us note here that for the purposes of object-based segmentation an explicit and accurate extraction of such information (e.g. shape or depth) carried in the apparent motion field is not necessary.

Such a motion-based approach is adopted in the first two of the three methods developed in this dissertation. The labeling is formulated as a grouping based on properties that are derived from the visual data and motion information plays the dominant role. A parametric model for the motion together with a spatial model of the label field provide the formal constraints for the definition of an object. Loosely speaking, the underlying assumption is that an object is an entity such that:

- Its kinematic characteristics [3] are described by a parametric model of relatively low order. In our case an affine model implies rigidity assumptions.
- Its localization[3] is compact.

In addition, the second method adopts a temporal model for the label field, that is, it implicitly imposes coherency in the way that the localization of each object changes over time.

One of the most important issues in the field of object-based segmentation is the trade-off between the generality of the assumed models and the degree of user interaction or initializations. From this perspective, the first two of the proposed methods

---

[3] We refer to the two-dimensional object projections on the image and not to the three-dimensional objects in the physical world

fall in the medium generality / limited interaction end (fig. 1.1). As far as generality is concerned, the global affine parametric models that are adopted by both can be too restrictive in the case of, for example, human motion. Furthermore, depending on the application and the user the goal of the segmentation may vary. The assumption that a decomposition of the scene based on the kinematic behavior of the depicted objects is possible might no longer hold true, thus other sources of information such as color and/or texture should be also used. The adoption of global models for such properties is in general more restrictive than in the case of motion. Furthermore, objects that are semantically meaningful might not exhibit a global homogeneity in their properties and/or discontinuities might be difficult to be detected.

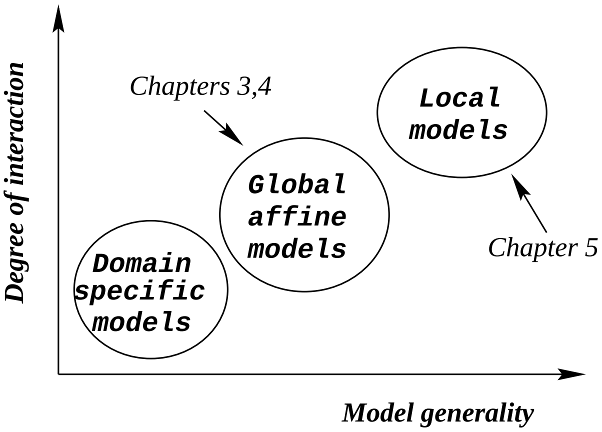


Figure 1.1: Tradeoff between generality and user interaction.

For all of these reasons the third of the proposed methods adopts more general models and transfers some of the difficulty of the problem to the user via an interaction phase. Loosely speaking, an object is considered as an entity which exhibits **local** homogeneity in its color and motion characteristics, it is locally compact in its localization and its localization changes coherently over time. In this method, the semantics about the contents of the scene are introduced in a user interaction phase in which the models of the objects are initialized. Conceptually, such an approach bears similarities with methods that train the parameters of a rather general model with data from a specific domain.

1.2 Major contributions and organization of the dissertation

The relation of our approaches to other methods in the field will become apparent in chapter 2, where we concisely review related works. Let us note here that common issues that arise in the area are on the one hand the reliability of the extracted properties on which the grouping is based and on the other, the localization accuracy of the labeling. In order to deal with both problems we used an initial partitioning of each frame into intensity (or color) segments in all of the three approaches. The partition is obtained using a method which operates on the gradient of the intensity or the color of each frame. Under the assumption that each of the resulting intensity(color) segments covers an area belonging to a single object, properties are extracted per intensity segment and the labeling is performed at intensity segment level. This offers the following advantages over methods that label each pixel separately:

Reduction of computational complexity The dimensionality of the problem is greatly reduced, since a much smaller number of sites need to be labeled.

Robustness Properties extracted from an intensity segment are in general more reliable than properties extracted from single pixels. Especially when motion is taken into consideration, more evidence exists for determining the kinematic behavior of an intensity segment than the behavior of a single pixel.

Localization accuracy Under the assumption that intensity(color) discontinuities are necessary conditions for object discontinuities, intensity(color)-based evidence is very useful because of its very good localization properties. An initial intensity(color) segmentation introduces such information in a very early stage.

By this scheme we define the following hierarchy, which we adopt throughout the dissertation. At the lower level there are *pixels*, at an intermediate level intensity(color) segments and at the higher level objects. The term *region* will be used rather loosely in the context of this dissertation to refer to a (usually) spatially connected set of pixels or intensity segments. For the remainder of the thesis, unless explicitly stated otherwise, the term *segmentation* will refer to the initial intensity(color) segmentation procedure while the term *labeling* will be used to refer to the process of assigning object labels to the intensity(color) segments. A pictorial representation of this hierarchy is given in fig. 1.2.

The remainder of the dissertation is organized as follows. In chapter 2 we concisely describe the proposed methods and position our methods in the related literature. In chapter 3 we present the first of the proposed methods. We propose a sequential approach in which the estimation of the kinematic behavior of the objects and the motion-based labeling of the intensity segments are performed in subsequent stages. The main contributions of this chapter can be summarized as follows:

- we develop a confidence measure for the motion field estimated by a Block Matching motion estimator. The Block Matching motion estimator is expressed in the probabilistic framework and the confidence measure is derived in terms of the *a posteriori* probability of the motion vector. The estimation of the confidence measure is incorporated in the estimation scheme of the Block Matching motion estimation and introduces negligible additional computational cost.

- we propose a robust clustering method which simultaneously estimates the parameters of a known number of models that describe the motion of the objects in the scene. Robust M-estimators and motion-specific confidence measures are

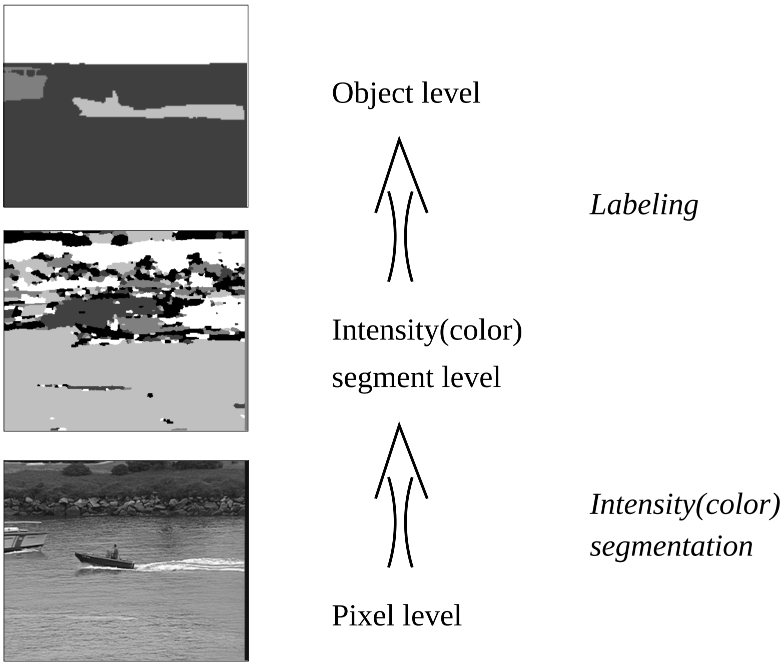


Figure 1.2: Hierarchy adopted by the proposed methods.

employed in the regression phase and intensity segments are used as primary elements.

- we model the label field as a Markov Random Field, where the sites are the intensity segments. In comparison with other approaches that use pixels as sites, this approach reduces the computational complexity and makes the labeling more robust.

In chapter 4 we present the second of the proposed methods. We propose an approach in which the kinematic behavior of the objects in the scene and the labeling of the intensity segments are jointly estimated. The main contributions of this chapter can be summarized as follows:

- we express the labeling problem in the Markov Random Field - Maximum *A posteriori* Probability framework in which the intensity segments that result from an independent intensity segmentation phase are used as sites. A number of iterative methods where the Markov Random Field is defined over pixel sites can be formulated as special cases of this method. The formulation addresses in a single framework the following issues:
  - imposition of spatial and temporal constraints on the label field
  - treatment of occlusions

  - utilization of intensity-based evidence to support the detection of motion discontinuities

- we propose an optimization method for the joint estimation of the parameters of the motion models and the label field that maximizes the *a posteriori* probability of the label field. The three-frame approach that was adopted introduces an intermediate directional field in the proposed iterative relaxation method.

In chapter 5 we present the third of the proposed methods. We propose a semi-automatic method in which the objects that the user outlines at the first frame of the image sequence in an interaction phase are modeled and tracked for the rest of the image sequence. The main contributions of this chapter can be summarized as follows:

- we propose a modeling of the local color and motion statistical properties of the objects. The region of support of each local model is adapted to the local characteristics of the independent color segmentation.

- we propose an optimization method for the maximization of the joint probability of the label field and the observed color and motion properties. The optimization is performed in terms of the statistical properties of the color segments.

In chapter 6 conclusions are drawn and open issues are discussed. In the same chapter we attempt a critical discussion on the design choices that we have made as well as on some of the alternative choices. Finally, appendix D gives a short description of the synthetic sequences that have been used.

Chapter 2

Related Work

In the previous chapter we gave an introduction to the field of object-based segmentation of image sequences. In this chapter we attempt a review of related works in the literature in order to position our work in the field. The chapter is organized as follows. In section 2.1 we give a short description of the proposed methods and the motivation behind the development of each one. Subsequently, we give an overview of the most related works organized in three sections, each one corresponding to one of the proposed methods. In each section a review of the related works from the prism of the corresponding method reveals its advantages and limitations.

More specifically, in section 2.2 we concisely review indirect motion-based segmentation methods, that is, methods for motion-based segmentation that depend on an independently estimated motion field. In section 2.3 we review methods for motion-based segmentation which, to a higher or lower extent, couple the motion estimation and the labeling problem. Finally, in section 2.4, we review works that allow objects with non-rigid motion patterns.

2.1 Overview of the Proposed Methods

The first two of the proposed methods attempt a motion-based decomposition of each frame of the sequence, under the assumptions that the number of objects is known and that the motion of each object can be described by an affine parametric model. This model imposes rigidity constraints on the kinematic behavior of the objects in the scene.

The first of the proposed methods falls in a general category of methods which rely on an independently estimated motion field. We propose the separation of the problem into two subsequent phases: the motion hypotheses extraction phase and the labeling phase. In the motion estimation phase, first a motion field is estimated using a hierarchical block matching motion-estimation scheme. A robust clustering technique is developed in order to estimate the parameters of the affine models that describe the motion of each object. The set of the estimated parameters serve as a set of motion hypotheses describing the kinematic behavior of the objects in the scene. In the labeling phase, an object label is assigned to each of the intensity segments. The labeling is expressed as an optimization problem under the assumptions that the motion-compensated intensity difference follows a Laplacian distribution and that the *a priori* probability of the label field is a Gibbs distribution. The latter implies that a Markov Random Field is defined over the graph whose nodes are the intensity segments that are extracted in the initial intensity segmentation phase. A deterministic iterative optimization method provides the final solution.

The motivation behind the development of such a method is to provide the first of the estimated label fields. In the motion estimation phase we address the issue of the simultaneous estimation of multiple motion hypotheses from an inaccurate motion field by the development of confidence measures and robust clustering methods. In the labeling phase, in contrast to most of the methods in the literature that depend on a pre-estimated motion field, we propose to use as evidence the motion-compensated intensity differences under the extracted motion hypotheses and not the motion residual. In this way we achieve much higher localization accuracy and with a three-frame approach we deal in a simple and efficient way with occlusions. Finally, the Markov Random Field modeling of the label field imposes spatial constraints which inhibit label fields with isolated intensity segments.

The second of the proposed methods attempts to solve jointly the motion hypotheses extraction and the labeling problem. The notion of Markov Random Fields (MRF) is used in order to express spatial and temporal constraints at the level of the initial intensity segments. The criterion is the maximization of the conditional *a posteriori* probability (MAP) of the label field given the motion hypotheses, the estimation of the label field in the previous frame and the image intensities. The equivalence between the Markov Random Fields and the Gibbs distribution is exploited and the labeling is formulated as an optimization problem with respect to the motion parameters and the label field itself. For the optimization we propose a method which reduces the corresponding objective function in an iterative way with respect to the motion parameters and the label field. In the optimization with respect to the motion parameters (*motion estimation phase*) the estimation is performed at segment level, that, is a set of parameters is estimated for the collection of intensity segments that comprise an object. The estimation is constrained by the intensity conservation principle and by the temporal coherency of the label field. In the optimization with respect to the label field (*labeling phase*) an object label is assigned to each of the intensity segments. The labeling is constrained by the intensity conservation principle and the spatial and temporal coherency of the label field. A three-frame approach is developed in order to deal with occlusions.

The motivation behind the second of the proposed methods is to couple in a single framework the interdependent problems of motion estimation and motion-based labeling. The motion estimation / motion-based labeling is formulated as a well-defined optimization problem where each component is modeled by a probability distribution. The proposed method addresses in a single framework classical issues raised in the field. More specifically the imposition of spatial and temporal constraints on the label field, treatment of motion occlusions and incorporation of intensity evidence for object boundaries. A number of iterative methods where the Markov Random Field is defined over pixel-sites can be formulated as special cases of the proposed method.

The third of the investigated methods attempts to deal with more complex objects that are not necessarily characterized only by their motion behavior. The goal is the development of a method capable of segmenting complex scenes into objects whose color and motion attributes vary. Since there is not a unique way of defining a segmentation into objects that are not homogeneous in their attributes (e.g. various motion and color patterns) user interaction is utilized. For the first frame of the sequence user scribbles are used to obtain the first label field and built a description of the local statistical properties of the objects. Subsequently, the label field is tracked for the rest of the sequence. The labeling is based on a probabilistic classification of the segments that result from the initial color segmentation scheme. It is assumed that the data at points inside the same color segment is generated by the same process; that process is modeled as a multivariate Gaussian. The conditional probability of motion and color given the label field, in a window around the center of each segment is modeled as a mixture of multivariate Gaussians, each one generated by a different object. The classification criterion is the maximization of the joint probability of the label field and the observations with respect to the label field. For the maximization of the joint probability a deterministic iterative local search algorithm is developed.

The motivation behind the third approach is to overcome the rigidity assumption of the previous two methods and deal with more complex scenes. A local modeling of the distribution and the *a priori* probability of each object can represent objects which exhibit variations in their motion, color and spatial characteristics. Furthermore, the assumption that the data at points inside the same color segment is generated by the same process allows to express the local object distributions as functions of the color segments' statistical representations and to perform the classification at color segment level.

2.2 Indirect Methods for Motion-Based Segmentation

Extensive work has been conducted in the field of motion-based segmentation of image sequences. From the prism of the first of the proposed methods we concentrate on approaches which depend on an independently estimated motion field. Depending on whether or not they utilize intensity-based evidence for the labeling we divide them in two main categories (fig. 2.1). Unless explicitly stated, all of the reviewed methods in this section assume that the motion patterns of the objects in the scene can be described by parametric motion models.

The first category consists of methods that attempt a segmentation of an independently estimated motion field disregarding intensity-based evidence of an object edge. Borshukov *et al.* [17] as well as Wang and Adelson [113] estimate the affine motion parameters of rectangular blocks and merge them according to a distance function defined in the parameter space. A threshold in the motion residual determines the reliable blocks and the merging procedure is terminated by a threshold in the distance function. The dominant motion estimation / outlier detection paradigm is adopted [4], where a threshold in the labeling phase determines which motion vectors follow the dominant-motion hypothesis. In a similar approach Wang and Adelson [113] merge the blocks applying a C-means algorithm in the parameter space. Both of the approaches depend

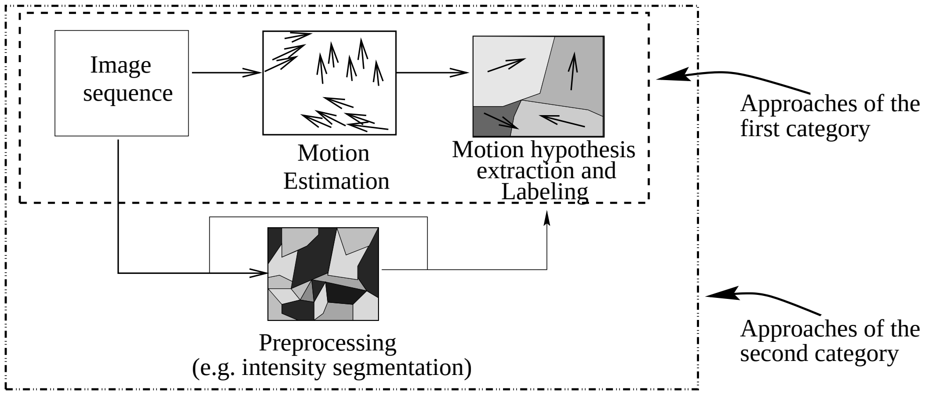


Image sequence
Motion Estimation
Motion hypothesis extraction and Labeling
Preprocessing (e.g. intensity segmentation)
Approaches of the first category
Approaches of the second category

Figure 2.1: Indirect methods for motion-based segmentation divided based on whether or not they utilize intensity information in the labeling phase.

on the parametric representations of each of the blocks, which can be quite sensitive to noise. Furthermore, the dependence of the method of Wang and Adelson [113] on the accuracy of the estimated motion field seems quite high since no attempt to increase the robustness of the C-Means algorithm is made. Nitsuwat and Jin [80] [81] robustify the classical C-Means algorithm by using fuzzy clustering techniques [28]. In [80] a clustering in parameter space by a fuzzy C-prototypes algorithm [36] is followed by a cluster merging phase. The latter is driven by a robust cluster similarity measure which is defined in terms of the degrees of membership of the motion vectors under the motion hypotheses that the cluster in question are associated with. The method exhibits the drawbacks of [113] as far as the parametric representations of blocks are considered and, furthermore, seems to produce rather “noisy” label fields. On the other hand, in principle, the use of robust clustering methods increases the accuracy of the estimated motion hypotheses.

Choi and Kim [24] propose a multistage segmentation of a dense optical flow field, where at each stage different motion-based merging criteria are applied. Under the assumption of a static camera a change detection map is estimated in a preprocessing phase. In a framework inspired by theories on evolution, Huang *et al.* [48] propose the use of a genetic partitioning algorithm for the segmentation of an optical flow field. The motion field is obtained by a point-correspondence algorithm. Each chromosome is a binary partition of the optical flow field such that it represents the region that supports one motion hypothesis. Crossover and mutation operators are trivially defined on the partitions. A *self-adaption* operator which corresponds to a combined robust motion estimation / classification step is used to re-estimate the region of support of the motion hypothesis associated with the chromosome. The ability of the genetic algorithm to simultaneously encode a number of motion hypotheses and the use of the binary partition maps seem to robustify the estimation of the parameters of the motion models. On the other hand, the method is computationally expensive if it is applied to a dense motion field. Finally, results are reported only in terms of the estimated motion parameters.

None of the above mentioned methods model the label field in order to enforce spatial constraints. In contrast, Murray and Buxton [74] model the label field as a Markov Random Field and express the labeling problem as an optimization problem in terms of the conditional *a posteriori* probability of the label field given an independently estimated optical flow field. Spatiotemporal cliques are defined and a stochastic relaxation method [53] is employed as an optimization mean. The Markov Random Field modeling introduces useful spatial and temporal constraints in the label field but the proposed optimization scheme is computationally intensive.

The methods reviewed so far propose a labeling depending solely on the information carried in an independently estimated motion field and therefore have a serious drawback: the accuracy of the labeling depends heavily on the accuracy of the estimated motion field. The introduction of spatial and temporal constraints [74] reduces the problem in areas within the objects but is of limited assistance at object borders. The reason is that at object borders a) the estimated motion field is highly unreliable and b) spatial constraints can be applied from either direction of the border. That is, a band of uncertainty exists at motion discontinuities and there the labeling is likely to fail. The width of the band depends on the motion estimation scheme and in methods that employ a block as a region of support it is roughly equal to the block size. In cases that the motion discontinuity is such that an occlusion is generated in the direction in which the motion field is estimated (e.g. backward) the width of such a band is roughly equal to the motion magnitude. In image sequences that contain large motion activity that is a serious drawback.

In order to increase the localization accuracy, a number of methods have been proposed that attempt to utilize intensity information in combination with an independently estimated motion field. In the multivalued mathematical morphology framework, Gu [45] adopts a hierarchical scheme in which at each level the analysis is performed on the motion residual between the modeling at the higher level and the motion field. Markers are declared at flat areas with respect to the motion residual and a watershed algorithm provides the label field at the current level. Affine parametric models are then used to model the motion residual for each of the resulting regions. In order to increase the localization accuracy, he introduces a final stage where he labels segments that result from an independent intensity segmentation phase with the dominant label in the intensity segment in question. Since a motion boundary usually separates two objects, the underlying assumption is that the majority of the motion vectors within the intensity segment in question are correctly labeled. Mansouri and Konrad [62] formulate the labeling problem as that of region competition and solve it using the level set methodology [99]. A number of parametric motion hypotheses are estimated by clustering a sparse motion field. A grouping that is based on the assumption that the distance between the features of the model-based projections of a pair of pixels should be very similar to the distance between the features of the pair of pixels themselves provides an initialization of the procedure. A C-means clustering refines the estimation of the motion parameters. The equations governing the evolution of the contours are defined in terms of the differences in the motion-compensated intensity differences that are generated by each of the competing motion hypotheses and the contour curvatures. Loosely speaking, the better the image intensities are explained by a motion hypothesis (compared to the competing motion hypotheses), the higher the speed for the evolution of the contour of the motion hypothesis in question. Although the C-Means algorithm is not robustified, the use of a more reliable sparse motion field and the fact that the clustering is not performed in the parameter space can lead to a rather reliable estimation of the motion hypotheses (subject to the initialization procedure). Finally, Altunbasak *et al.* [3] extract the motion hypotheses by C-Means clustering of a dense motion field. This is followed by a procedure which iterates between a classification and an estimation phase. In the classification phase color segments are labeled according to the motion-compensated intensity difference. In the estimation phase the motion parameters are re-estimated based on the labeling. Although the iterative procedure is supposed to improve the estimation of the motion parameters, convergence is not guaranteed since the two stages minimize different objective functions. A threshold in the motion-compensated intensity difference is used to reject unreliable motion vectors.

Based on the above classification the first of the proposed methods is most related to the second category of approaches. Like Mansouri and Konrad [62] we claim that the motion parameters of multiple motion hypotheses can be accurately extracted from a pre-estimated motion field. Our motion hypothesis extraction is similar to the C-means clustering of Altunbasak *et al.* [3]. However, we incorporate robust estimation methods in the clustering procedure (e.g. [80]) and we drastically reduce the computational complexity by estimating matrices that give in closed form the mean-square motion residual of an intensity segment. Furthermore, we propose a new confidence measure which has a probabilistic interpretation given the motion estimator that was used and does not need any user-defined parameter. We propose a Markov Random Field modeling of the label field that combines the advantages of [74] and [45] by adopting intensity segments as sites. Finally, in contrast to all reviewed methods, we use a three-frame approach in order to deal with occlusions.

2.3 Direct Methods for Motion-Based Segmentation

Motion information is one of the main elements that are used for segmenting video sequences. However, extracting and coupling motion information with the labeling process is by no means a trivial task [109]. For the estimation of motion, spatial constraints need to be imposed in a form of a support region where the motion is assumed either to be smooth or to follow a parametric model. In general, if the region of support is arbitrarily chosen then the motion estimate will deteriorate either because the single motion assumption within the region is violated or because the texture pattern is too low to constrain sufficiently the estimation. From the labeling point of view inaccurately estimated motion information leads to inaccurate labeling results. Furthermore, in the motion-based segmentation framework issues like the occlusions and the temporal coherency of the label field need to be addressed. The former concerns areas that appear or disappear from the scene due to motion. Hence, information about their temporal behavior is limited or even absent. The latter provides useful constraints between the kinematic behavior of the objects and their localization in consecutive frames.

A number of methods for performing motion-based segmentation have been proposed over the last decade. In the framework of the second of the proposed methods we will concentrate on five categories of approaches (fig. 2.2). To the first category belong methods which simultaneously estimate the motion information and its region of support. Depending on if the label field is explicitly defined, temporal and spatial constraints are imposed either on motion and/or on the label field itself. In the following four categories we classify methods that combine an initial intensity segmentation with motion information. First, top-down approaches which are based on the dominant-motion-estimation / outlier-detection paradigm are described. Second, we review methods in which a region-merging process is driven by motion-based distance measures. Third, we present methods that utilize an initial intensity segmentation in order to incorporate spatial constraints in the Expectation Maximization framework. Finally, methods that combine the Markov Random Field modeling with an initial segmentation are presented.

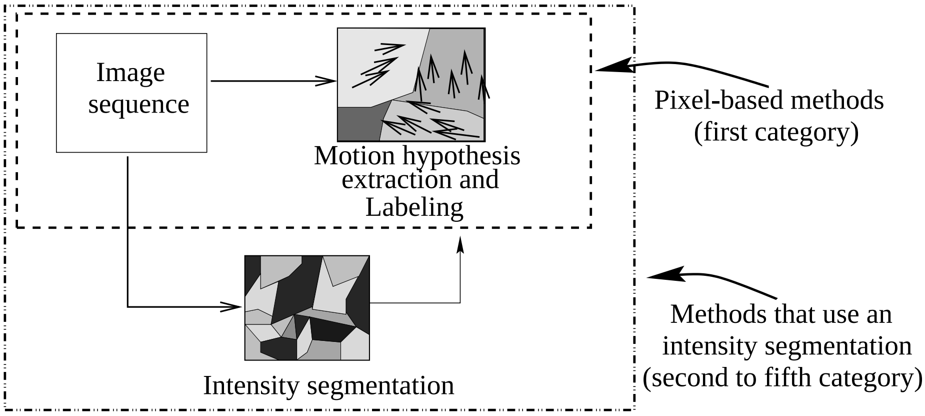


Image sequence
Motion hypothesis extraction and Labeling
Pixel-based methods (first category)
Intensity segmentation
Methods that use an intensity segmentation (second to fifth category)

Figure 2.2: Direct methods for motion-based segmentation.

In the first category we begin with methods that attempt to overcome the isotropic smoothing that the Horn-Schunck algorithm [47] imposes. These approaches are oriented to the regularization of the motion estimation in such a way that discontinuities in the motion field are preserved. Black and Anandan [14] propose the use of robust statistics for the detection of outliers both in the optical flow constraint and in the regularization term in an attempt to obtain piecewise smooth motion fields. In their work the discontinuities are expressed in terms of the motion field itself, while Nagel [76][77] introduces an *oriented smoothness constraint*, which suppresses the smoothing in the direction of the local intensity gradient. In other works a *line process* [42] is used in order to explicitly model the motion discontinuities. Konrad and Dubois [55] model the motion and the discontinuity fields as a pair of coupled MRFs and minimize the resulting energy function by means of stochastic relaxation. Identifying the need to exploit intensity discontinuities to detect motion discontinuities, they propose a potential function for the line field which depends on the local image gradient. However, in all of the above-mentioned methods the aim is the estimation of motion information rather than the motion-based segmentation. A label field is not explicitly defined and should be extracted in a later step from the discontinuities in the motion field. Furthermore, temporal constraints are not addressed in any of these methods. Stiller [104] explicitly models the label field, the occlusion and the motion fields as MRFs. In this way spatial and temporal constraints are introduced. However he still uses motion discontinuities as the means for detecting the borders of the objects and does not constrain the motion field with parametric models. Furthermore the cliques in the MRF formulation are defined over the pixels which results in a computationally intensive optimization process. An iterative scheme for motion estimation and labeling is proposed by Chang *et al.* [23]. The label and the motion fields are modeled as MRFs and the dense motion field is additionally constrained by multiple parametric models. Temporal coherency issues are not addressed however, and the MRF cliques are still defined at pixel level. Bouthemy and Francois [18] in the Markovian framework propose a method for motion-based segmentation of image sequences based on parametric motion fields. The cliques are defined over 2x2 blocks and not over pixels. However, such an arbitrary initial decomposition might violate object borders and provides robustness only to such an extent that the image structure in the block disambiguates the motion constraints. Furthermore, the neighborhood structure for the definition of cliques is not affected in comparison to pixel-based approaches; in both cases they are defined on a regular lattice. Finally, the temporal constraints are not incorporated in the optimization procedure but are merely used in the initialization phase.

A number of methods have been developed which utilize an initial intensity based partition to constrain the label field (fig. 2.3). In this framework hierarchical approaches [35][31] identify independently moving objects as collections of segments that do not conform to the estimated parametric dominant motion. Diehl [31] validates the dominant motion hypothesis independently per segment while Fablet *et al.*[35] explicitly express the spatial interactions between the segments by a MRF modeling of the label field. These top-down approaches are faced with the problem of estimating the dominant motion in the presence of multiple independent motion patterns. Furthermore, they impose an artificial hierarchy in determining the motion characteristics of the objects and may lead to situations where outlier segments do not belong to any object [67]. On the other hand the Markov Random Field modeling which Fablet *et al.* propose, although it does not address the temporal consistency of the label field, once presented with a reliable motion estimation can make use of a very fine initial partition and accurately recover the motion boundaries.

In the third category of approaches a parametric motion model is estimated on segment basis and the motion parameters are subsequently used to group the segments into regions with coherent motion behavior. Dufaux *et al.* [34] estimate the motion of each segment using a matching technique [68] which searches directly in the parameter space. A k-medoid clustering [51] in the parameter space groups segments with similar motion. A clustering approach in the parameter space is also proposed by Wang and Adelson [113] but that approach is sensitive to the errors in the parametric representation [2]. Moscheni *et al.* [67] and Wang [112] perform region merging and dynamically

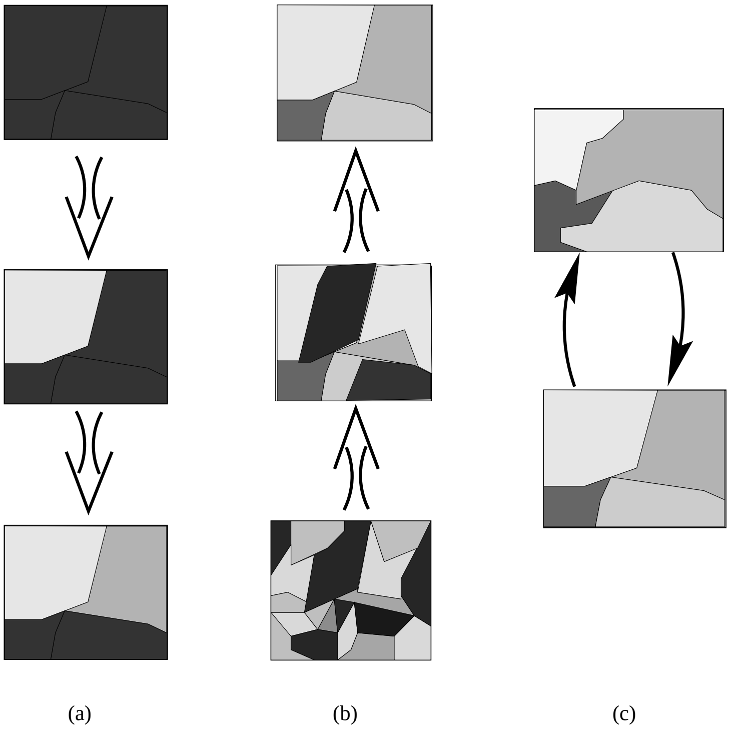


Figure 2.3: Three approaches to the motion-based segmentation problem: (a) the top-down approach (second category), (b) the bottom-up approach (third category) and (c) the region-competition approach (fourth and fifth category)

update the corresponding Region Adjacency Graph (RAG). The intensity segmentation is merely used for initialization of regions. Moscheni defines spatiotemporal similarity measures based on statistical tests between the nodes of the RAG, while Wang defines a motion-based distance which relies on the motion-compensated error before and after the merging. However, in both methods the merging is irreversible and the initial parametric motion estimations, which are performed independently per segment, can be unreliable depending on the size and the local image structure in the segments. Moscheni uses the same motion estimation scheme as in [68], which involves compensation of the camera motion, while Wang performs a Least-Squares regression on an independently estimated block-based motion field.

In the fourth category of approaches belong methods that utilize spatial and/or temporal constraints in the EM framework. Brady and O'Connor [19] in the *Expectation* step estimate the conditional probabilities of the label of each pixel by adjusting the *a priori* probabilities of the labels in a so-called contextual step. These contextually adjusted priors are estimated according to spatial constraints derived from an initial intensity segmentation and according to temporal constraints derived from a prediction of the label field. Their formulation favors the labeling with the same label of pixels within the same intensity segment, but no spatial constraints are applied between neighboring segments. In the *Maximization* step a hard classification is assumed, that is each pixel is considered to be labeled with a single “object” label. Weiss and Adelson [115] assume that the pixels within the same intensity segment are generated by the same process and estimate the *a posteriori* probabilities in the *Expectation* step by summing the deviations from the model prediction for all pixels within the same intensity segment. Their formulation addresses neither temporal coherency issues nor spatial constraints between pixels in neighboring segments. Furthermore, both of the works adopt a two-frame approach and do not address issues related to occlusions.

Finally, there are a number of methods which combine the MRF modeling of the label field with an initial intensity segmentation. Konrad and Dang [54] define cliques at pixel level but in the optimization procedure mergings on segment level are considered. Since in the MRF energy formulation no distinction is made between segments and regions, the initial intensity segmentation is used only as an initialization of the regions. Furthermore, in the optimization the mergings are irreversible. Gelgon and Bouthemy [40] propose a modeling using Markov Random Fields by considering cliques between adjacent segments. The potential of a two-segment clique is a function of the discrepancy between their parametric motion fields and a geometrical “compacity factor”, which is a function of the length of the common border and the distance between their centers of gravity. Although the discrepancy between the parametric fields is a better dissimilarity measure than the distance in the parameter space, the method depends on the assumption that the local image structure within each segment sufficiently constrains a reliable motion estimation. In [41] an extension of their method is presented, in which a term which favors a low number of labels is added to the formulation. In both of [40] and [41] the motion estimation is performed independently per segment and the temporal constraints are introduced only in an initialization phase in which the label field of the current frame is predicted from the label field of the previous frame. However, in [41], the motion parameters that are used for the initialization of the label field are estimated for each of the collections of segments that constitute an object.

In terms of the above classification the second of the proposed methods is mostly related to the methods of the first and fifth category in the cases that the label field is explicitly modeled as a MRF. Our work can be regarded as an extension of methods that define cliques at pixel level in the Markovian framework and adopt an optimization procedure which jointly estimates the motion and the segmentation field. We exploit the ability of such approaches to incorporate the spatial and temporal constraints in the optimization procedure. However, by defining cliques on segment level we provide tighter constraints for the labeling and reduce the dimensionality of the problem. The initial intensity segmentation groups together pixels in which the low degree of texture implies inadequate information about their temporal behavior. These segments are more reliable entities than pixels when used as primary elements for the labeling problem. The relation of our approach to existing pixel-based methods will become more apparent once the modeling and the optimization procedure have been described. In order to make this relation more clear, in appendix 4 we will present the degenerate case where the segments that result from the initial segmentation contain a single pixel. We will show that in that case the second of the proposed methods reduces to an approach that falls in the first category.

Finally, of all the methods that utilize an initial intensity segmentation, the work of Gelgon [35] [41] [40] is related most to our proposal. However, his way of combining the motion information with the labeling is quite different. The dominant motion estimation / outlier detection paradigm which is adopted in [35] has the shortcomings of the hierarchical approaches. In [41] and [40] motion is estimated independently per segment. Estimating motion parameters per segment requires sufficient local intensity structure, which often implies that the size of segment should be rather large. In search for sufficient texture the initial intensity segmentation method might violate significant borders. In our approach a region-based motion estimation is employed. The motion estimation is constrained by the intensity pattern of the whole region and by the temporal coherency of the label field. Therefore it is not crucial if some of the segments do not provide sufficient constraints. The ensemble of the constraints in the whole region is what determines the accuracy of the motion estimation. Furthermore, in both [41] and [40] the temporal constraints are introduced only in the initialization phase for the prediction of the initial label field. In comparison our approach incorporates the temporal constraints in the optimization procedure itself.

2.4 Tracking of non-rigid objects

From the prism of the third of the proposed methods we will concisely review methods that allow objects with non-rigid motion patterns. We will concentrate on three main categories of methods (fig. 2.4). To the first category belong methods that track a non parametric contour of the object. To the second category belong methods which attempt a parametric contour tracking. Finally to the third category belong methods which model statistically the properties of the regions that correspond to each of the objects that are present in the scene.

To the first category of approaches we classify the works of Nguyen and Worring [79] and Paragios and Deriche [89]. The latter employs a geodesic contour model [21] in order to track objects in a static background. Their approach seems to exhibit good localization properties and robustness, but it is not straightforward how to overcome the assumption of a static camera. Nguyen and Worring adopt an approach in which the localization of the contour for the current frame is driven by a minimization of an energy generated from three different sources; a motion-based prediction of the contour based on the estimated contour in the previous frame, an edge map estimated for the current frame and (possibly) an internal energy term imposing contour smoothness. In order to deal with clutter, they suppress the background edges. The latter are detected by examining whether their motion follow the parametric motion model that is estimated for the tracked object. Their method seems to track very well small and fast-moving objects, but might not be so robust in presence of non-rigid motion. The reason for the latter is that a parametric motion model is used for the contour prediction and for the suppression of the background edges.

A number of tracking algorithms are developed based on parametrized representa-

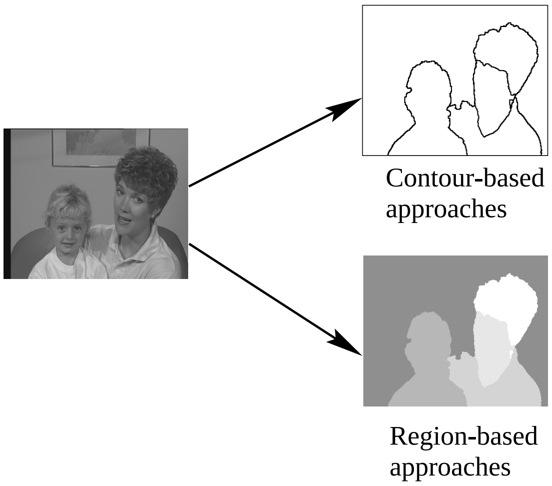


Figure 2.4: Contour-based (first and second category) and region-based (third category) methods for objects with non-rigid kinematic behavior.

tions of the shape and/or of the motion of the tracked objects [49] [106] [94] [16]. A parametric representation of the shape has the advantage that the tracking can be expressed as an estimation problem with respect to the parameters of the representation. In that framework temporal stability can be imposed by the adoption of a model for the evolution of the states associated with the parametric representation. Furthermore in the case that *a priori* knowledge of the domain is available [59], robustness and accuracy can be increased by the adoption of specific models (e.g. for humans [50] or hands [95]). On the other hand, the adopted models usually capture global shape properties and might be sensitive to changes in the contour topology.

Finally, in the third category of approaches belong methods in which the label field is determined based on the underlying distribution that describes the data of each object [85] [22] [82] [107]. Such works originate from research conducted in the area of stochastic model-based image segmentation (e.g. [120] [87] [42] [29]). Usually, parametric probability density functions are assumed whose parameters are either obtained from training sets or, in more elaborate schemes, are estimated in an iterative segmentation-estimation scheme.

Paragios and Tziritas [90] and Ayer [6] are examples of methods in which the observed data is modeled as a mixture of unimodal distributions (e.g. [107] [115] [116] [19]). Paragios and Tziritas compensate for the camera motion with a dominant motion estimator and model the displaced frame difference as a mixture of two Laplacians corresponding to the static and to the mobile assumption, respectively. Spatial constraints are imposed by modeling the label field as a Markov Random Field. Ayer assumes parametric motion models for a known number of objects and adopts a Gaussian model for the distribution of the motion-compensated intensity differences for each object. In the *Expectation Maximization* framework he treats the motion-compensated intensity differences as the **observed** data and the label field as the **hidden** data. Such approaches based on the statistics of the motion-compensated intensity differences implicitly or explicitly require a global motion model, either for the background and/or for the objects present in the scene. In the former case, the assumption that a number of objects move in front of a static (or rigidly moving background) may be violated (or the dominant motion estimation may fail). In the latter case, a global parametric motion model imposes rigidity assumptions.

In the mixture decomposition framework a number of methods have been developed in which multivariate multimodal distributions have been used in order to model the statistical properties of complex objects [85] [22] [82]. Oliver *et al.* model with a mixture of multivariate Gaussians the color (in normalized RGB space) and position properties of the face and the background in a facial-expression recognition system. Prior knowledge is incorporated by estimating the parameters of the "face Gaussians" from a training set. The derived model adaptively modifies its parameters by an online version of the *Expectation Maximization* algorithm at each frame. The label field is obtained using the maximum *a posteriori* probability criterion, once the parameters of the models are estimated. Chalom also utilizes the *Expectation Maximization* algorithm in the statistical modeling framework. In his work the motion and color properties of each object are modeled as a mixture of multivariate Gaussians. The number of the modes and the parameters of the Gaussians for the first frame are obtained by applying the *Expectation Maximization* algorithm on training data obtained by a user interaction phase. The training data for each of the objects are provided at pixels marked by a user-defined scribble. Once the parameters of the Gaussians are estimated a classification of each pixel according to the Maximum Likelihood principle provides the label field for the current frame. A motion-based projection of the "training-scribble" of the current frame into the next frame, provides the "training-scribble" on which the mixture parameters for the object in question will be estimated in the next frame. In an extension of the work of Chalom, O'Connor *et al.* [82] update the stochastic model of each object in the scene by iteratively applying the *Expectation Maximization* algorithm at each frame. In their work no temporal tracking was performed.

In terms of the above classification, our method is closer to the approaches of the third category and in particular the works of Oliver *et al.* [85] and Chalom [22]. In both a mixture model is used for representing the statistical properties of the objects. Since *a priori* knowledge, and therefore training data, of the domain is not available ([85]) we adopt a user interaction phase with scribbles as in [22]. In contrast to [22] we do not estimate the statistical properties of the objects only on the user scribbles. Instead, we use the user scribbles as markers in a color watershed segmentation [65] [64] [100] and the statistical properties of the objects are estimated on the resulting label field. Furthermore, similarly to [82], we iteratively update the objects' models iteratively at each frame by minimizing a well-defined probabilistic measure. In contrast to [82] we address the temporal aspect of the label field by defining an initialization of the objects' models based on a motion-compensated projection. Finally, in our approach the labeling is based on a probabilistic classification of *segments* that result from an initial color segmentation and not of pixels as in all above-mentioned methods. We assume that the data at points inside the same "color segment" is generated by the same process. This allows us to express the local object distributions as functions of the color segments' statistical representations and perform the classification at a segment level. This reduces the computational cost and introduces spatial constraints that inhibit noisy label fields with isolated pixels. Furthermore, a classification at segment level is more robust than that at pixel level, since features like motion and color characteristics are more reliably estimated for a color segment than for a single pixel. This holds especially for motion where erroneous estimations at problematic areas are quite common.

Chapter 3

Sequential Motion Estimation and Segmentation

In this chapter we present a method for motion-based segmentation of image sequences under the assumption that the number of the objects is known and that the motion of each object can be described by an affine parametric model[1]. A two-stage approach is adopted, where the motion estimation and the labeling are performed in subsequent stages (fig. 3.1). In the first stage a set of motion hypotheses is extracted and in the second stage segments that are extracted in an intensity segmentation phase are labeled according to motion-based properties and a model of the label field.

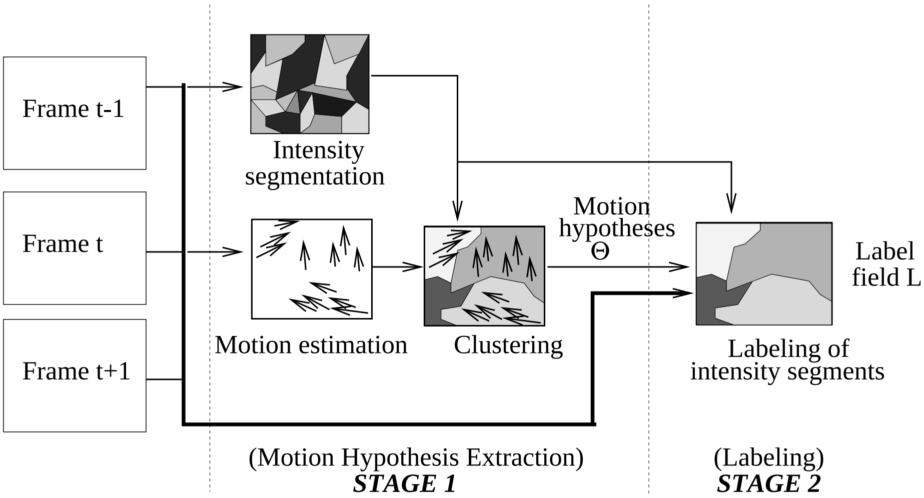


Figure 3.1: Outline of the sequential approach

---

[1] A preliminary version of this work appears in [91]

More specifically, in the motion hypotheses extraction stage (**Stage 1** in fig. 3.1) a motion field is estimated using a hierarchical block-matching algorithm. A clustering into a known number of affine models provides the hypotheses of the kinematic behavior of the objects in the scene. The goal is the accurate extraction of the parameters of the affine motion models given that the estimated motion field is bound to be inaccurate. In order to deal with the inaccuracies in the dense motion field, we developed a clustering method which uses robust statistics and motion-specific confidence measures.

In the labeling phase (**Stage 2** in fig. 3.1) each of the segments that result from the intensity segmentation phase is given an object label according to the motion hypothesis to which it conforms. The property used to quantify the degree of conformity of the motion of an intensity segment with each of the extracted motion hypotheses is based on the motion-compensated intensity differences defined over the pixels of the segment. A zero-mean Laplacian distribution is used to model the motion-compensated intensity differences for each object. Furthermore, a model of the label field itself is developed in order to enforce its spatial homogeneity. To encourage spatial homogeneity we model the label field as a Markov Random Field where cliques are defined as pairs of neighboring intensity segments. With this modeling the labeling problem is transformed in a well-defined optimization problem, where the objective is the maximization of the conditional *a posteriori* probability of the label field given the motion hypotheses. An iterative deterministic relaxation algorithm is used to solve the optimization problem.

The remainder of the chapter is organized as follows. In section 3.1 we describe the initial intensity segmentation algorithm and in section 3.2 we present the motion hypotheses extraction stage. Section 3.3 describes modeling the conditional *a posteriori* probability of the label field given the motion hypotheses. Finally, in section 3.4 conclusions are drawn.

3.1 Intensity segmentation

At the lower level of the proposed method (fig. 3.1) an intensity segmentation algorithm is applied on the current frame. We aim for conservative partitioning of the current frame, such that significant object boundaries are not violated. That is, we favor oversegmentation since the proposed method is not able to recover from initial undersegmentation by splitting a segment that does not entirely belong to a single object. Although the choice of the intensity segmentation method is not restrictive to the generality of our approach we favor methods which consider the intensity gradient rather than clustering approaches. For its low computational complexity and good edge localization accuracy we use the watershed segmentation algorithm [12]. A filtering with morphological operators [101] with a small (3×3) structuring element is used for a nonlinear smoothing of the current frame. Once the noise level is reduced the morphological gradient is estimated and segment *markers* are extracted as areas where the gradient is lower than a threshold (fig. 3.2). The flooding procedure described by Vincent [110] provides the final partition.

The threshold for the marker extraction is a user-specified prediction of the upper threshold for the gradient within each object. The underlying assumption is that within

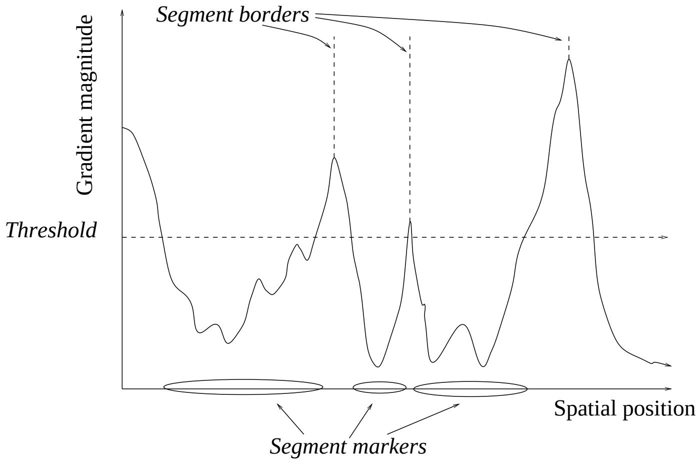


Figure 3.2: Initial intensity segmentation in the 1-D case

each significant object exists an area that is smooth enough so that the gradient magnitude is lower than the threshold. Edges with gradient magnitude smaller than the threshold are not preserved. It should be noted that the threshold is not directly related with the amount of texture within a segment. During the flooding procedure a segment will encapsulate some of the pixels which lie between its marker and the marker of the neighboring segment and have a higher gradient magnitude than the threshold (fig. 3.2).

The proposed intensity segmentation method has been used throughout the thesis and experimental results are presented in the corresponding sections. In general, it exhibited good localization accuracy for a wide variety of image sequences. Here we will present results for the image sequences "train" and "sunflower garden", the latter serving as an example of its limitations. The original frame and the corresponding intensity segmentation for the "train" sequence are depicted in fig. 3.3. This figure illustrates a typical initial segmentation obtained for the image sequences used in this dissertation. Although the localization accuracy cannot be easily observed due to the severe oversegmentation, it will become apparent when the intensity segments are labeled according to their motion characteristics.

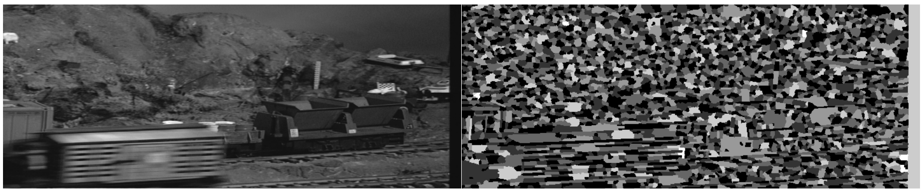


Figure 3.3: 10th frame and the watershed segmentation for the "train" sequence

As any intensity segmentation algorithm, the proposed method cannot guarantee that all of the object edges will be preserved. Besides the obvious case where the object edge does not coincide with an intensity edge, the proposed method also fails for very thin and elongated objects. The reason is that, although small, the morphological operators that are used for the estimation of the gradient indicate high gradient magnitude and therefore no marker is set within such objects. An example is shown in fig. 3.4, where we present the 10th frame and the corresponding intensity segmentation for the "sunflower garden" sequence. The method operates well for the trunk and the large branches of the tree but, as expected, merges the thin branches with the sky.

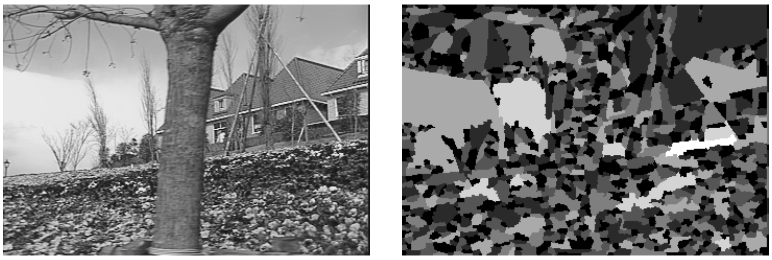


Figure 3.4: 10th frame and the watershed segmentation for the "sunflower garden" sequence

Once the intensity segments are extracted a Region Adjacency Graph (RAG) can be built to express the neighborhood relations between them. Let us define the mathematical notation that will be used for the rest of the thesis. We denote with $\{s : s \in [1 \ldots K]\}$ the set of watershed segments, with G_s the set of pixels in the watershed segment s and with $R = (V, E)$ the corresponding RAG. The set of nodes V is the set of watershed segments and E the set of edges that connect neighboring segments. Let us also denote with N_s the set of neighbors of segment s.

$$N_s = \{s' : (s, s') \in E\} \tag{3.1}$$

3.2 Motion Hypothesis Extraction

In this stage, the goal is the extraction of a set of hypotheses about the kinematic behaviors of the objects in the scene. In order to do so we need to address the following issues

1. What is a *"motion hypothesis"*.
2. Which features are used.
3. What is the estimation technique.

The first issue is related to our assumptions about the kinematic behavior of each object and typically it includes the type of the motion model that is used. The higher the order of the model, the more complex the object motions that it can describe but also the less robust the estimation of its parameters. Furthermore, the higher the order of the model, the higher the dimensionality of the solution space (i.e. motion parameters and label field). Given that there is inevitable noise in the observations based on which the motion parameters will be estimated, higher dimensionality also implies a larger number of spurious local minima. A good compromise between robustness and complexity which is widely adopted in the motion-based segmentation and tracking framework is the affine model. Such a model is also adopted in the context of the method developed in this chapter as sufficient to describe the 2D motion field induced by each object. Such a model, described in eq. 3.2, is linear with respect to the motion parameters and describes rigid 2D motions with translational, rotational and skew components. This corresponds to 3-D affine motion of planar surfaces under an orthographic camera model [105]. More specifically, let us denote with $\mathbf{i} = (\mathbf{i}_x, \mathbf{i}_y)$ a pixel and with $\theta = \{\theta(1), \ldots, \theta(6)\}$ the affine parameters that describe the kinematic behavior of the object to which the pixel $\mathbf{i}$ belongs. Then, the model-generated motion vector $\tilde{\mathbf{v}}_{\mathbf{i}}$ at pixel $\mathbf{i}$ is given by the following equation:

$$\tilde{\mathbf{v}}_{\mathbf{i}} = \left[\begin{array}{c} \theta(1)\mathbf{i}_x + \theta(2)\mathbf{i}_y + \theta(3) \\ \theta(4)\mathbf{i}_x + \theta(5)\mathbf{i}_y + \theta(6) \end{array} \right] \tag{3.2}$$

The second issue is related to the kind of evidence that is used in order to estimate the model parameters. The methods in the literature we can be divided in the following two groups:

Indirect methods , which estimate the parameters of the model based on an independently estimated motion field (e.g. [1], [74], [113]).

Direct methods , which estimate the parameters of the model by considering such "raw" features as the image intensities or the intensity derivatives (e.g. [83], [14]).

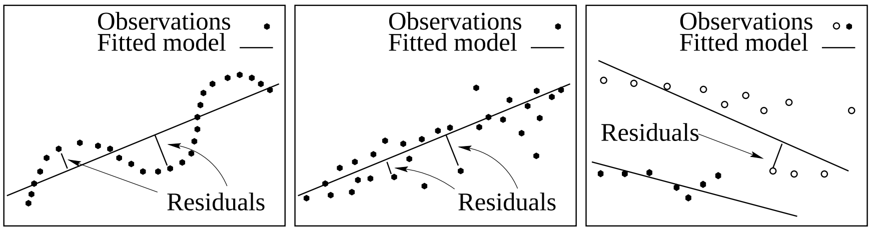
An in-depth discussion about the differences of the direct and indirect approaches goes beyond the scope of this thesis. Let us only note that the intensity-based evidence can provide higher accuracy but is in general more susceptible to local minima. In order to overcome the problem, direct methods are usually incorporated in multiscale schemes. In the context of the method described in this chapter we adopt an **indirect** approach and in particular we adopt the block-matching estimation scheme. Despite the inefficiencies of the latter which are related to the fixed block size, fixed accuracy, computational complexity and (usually) absence of spatial constraints it is an estimator whose variants have been widely adopted in coding schemes. The reason is the simplicity in its concept and the high level of potential parallelism which makes it particularly suited for hardware implementations.

The questions that the third issue raises are related to the estimation of the parameters of multiple models in the presence of noisy measurements. This includes the definition of a quantitative measure of how well an observation (e.g. a motion vector) conforms to a given model. Typically, the definition of such a quantitative measurement expresses indirect assumptions about the type of the residual distribution (e.g. Gaussian, Laplacian, etc.). The residuals, that is, the discrepancies between an observation and the underlying model, originate from three different sources:

Violations of the model assumptions Such a situation occurs if the model is not sufficient (e.g. it is of too low order) to describe the observations. An example is given in fig. 3.5(a) where a linear model is employed to describe the observations generated by a sine function.

Inaccurate observations Such a situation occurs if, due to noise or failures in the measurement procedure, the observations are not accurate. In the case of indirect methods, residuals of this kind are produced due to the errors that are introduced by the motion estimation method. An example is given in fig. 3.5(b) where we depict the residuals of noise-contaminated observations generated from a linear model and the linear model itself.

Misclassifications Residuals of this kind are connected to the clustering procedure and are caused by misclassifications of a number of observations. An example is given in fig. 3.5(c) where the observations are generated from two linear models.



(a) Residuals due to model insufficiency

(b) Residuals due to inaccurate observations

(c) Residuals due to misclassifications

Figure 3.5: Types of motion residuals

A distinction between the three different types of residuals is not trivial since a residual usually originates from a combination of more than one source. Often, estimation schemes assume that the residuals follow simple distributions such as Gaussians. Such schemes are sensitive to the presence of large residuals of the second and third category. In order to deal with it, we develop a clustering method inspired from robust statistics and the C-Means algorithm [33]. In comparison to the original C-Means algorithm it has the following characteristics:

- An affine model is assumed to describe the data (i.e. motion cluster) within each cluster. In comparison, the original C-Means algorithm considers a translational motion model.
- Motion-specific confidence measures are incorporated in the clustering. Such measures attempt to identify wrongly estimated motion vectors and diminish their influence in the regression scheme. This is connected to residuals of the second category.

- A robust regression method is employed for the estimation of the parameters of the models. This aims at decreasing the effect of large residuals in the parameter estimation. Such large residuals typically belong to the last two categories or to their combination.

The remainder of the motion hypotheses extraction section is organized as follows. In subsection 3.2.1 we will discuss issues related to the confidence measures and in subsection 3.2.2 we will describe the robust regression/clustering scheme. Experimental results will be presented in each of the two subsections.

3.2.1 Confidence measures

Motion estimators are known to be prone to errors originating from a variety of sources such as occlusion phenomena, absence of texture and support region that strives over motion discontinuities. Confidence measures have been developed by the realization that it is possible to identify such sources of errors and subsequently quantify the reliability of the estimation of a motion vector. The usage of such a quantification measure is two-fold. On the one hand, it can be incorporated in the estimation procedure itself in order to increase the estimation accuracy [43] [117]. On the other hand, it can provide useful information to the process in which the motion field is intended to be used. Lundmark *et al.* [61] use motion vector certainty in order to reduce the bit rate in video coding and Altunbasak *et al.* [3] in order to discard unreliable motion vectors in a regression scheme.

A number of methods have been developed in order to express the degree of confidence in an estimated motion vector (e.g. [30] [102]). Most of them are developed having in mind motion estimation techniques that rely on the optical flow constraint and lead to confidence measures in terms of the eigenvalues of the covariance matrix of the spatial derivatives of the intensity. Barron *et al.* [7] suggest using of the minimum eigenvalue as a measure of reliability, while Ghosal and Vanek [43] suggest the sum of the eigenvalues. While the first utilize the confidence measure in order to evaluate the accuracy of the motion estimator at different densities of the motion field, the latter incorporate it in the estimation scheme and impose anisotropic smoothness constraints. Simoncelli *et al.* [103] formulate the motion estimation problem in a probabilistic framework in order to derive probabilistic distributions of the motion. Although his approach is not primarily aimed at the derivation of a scalar confidence measure, he computes one as the trace of the covariance matrix of the *a posteriori* probability of the motion vector. The latter is derived by modeling the prior distribution of the motion vector, the noise in the estimation of the derivatives and the discrepancy between the "true" motion field and the apparent motion field. Dev *et al.* [30] arrive at a similar measure by performing an error analysis of the assumed image motion model. In another similar approach, Yoshida *et al.* [117] propose a measure which quantifies the sensitivity of the block-based motion estimator in certain directions. The measure is defined in terms of the spatial intensity derivatives and is used to merge blocks with the same directional sensitivity in order to re-estimate more reliably their motion. His method does not seem to estimate the reliability of the **estimated** motion vector but rather the expected sensitivity of the estimation in various directions. Finally, for the block matching estimator Lundmark *et al.* [61] use as confidence measure the weighted sum of the motion-compensated intensity differences in the block in question. Such measure does not take into consideration the statistics of the motion-compensated intensity differences in for example flat areas and does not adapt to the different image sequences.

Our motivation for the development of a confidence measure is to reduce the effect of wrongly estimated motion vectors in a regression/clustering scheme where the motion parameters will be extracted. To the best of our knowledge, all current schemes that are faced with such a problem discard motion vectors whose either confidence is higher than a threshold or their rank based on the confidence measure is higher than a threshold. Such an approach has the disadvantage that a method has to be developed for adapting the threshold according to the characteristics of the scene. This becomes especially difficult if the design of the confidence measure does not allow a meaningful relative comparison between two confidence measures. An example of such a confidence measure is the motion-compensated intensity difference and a counterexample the measure derived by Dev *et al.* [30]. Furthermore, with the exception of [117] and [61], the above mentioned confidence measures were developed for motion estimators that are based on the optical flow constraint. Although there are a lot of common elements between the various estimation schemes, the errors depend on the particular estimator that is employed. Here, we will concentrate on a hierarchical block-matching motion estimation scheme and derive confidence measures that express the confidence in the estimated motion field. Like Simoncelli, we formulate the motion estimation problem in the probabilistic framework and derive the confidence measures in terms of the *a posteriori* probability of the estimated motion vectors at each level of the multiscale scheme. In contrast to him we do not make assumptions about the prior distribution of the motion vector and estimate the *a posteriori* probability of the motion vector by an estimation of the prior distribution of the intensity which is provided as a by-product of the search scheme. The type of the conditional probability distribution of the motion-compensated intensity differences is derived from the objective criterion of the block-matching motion estimator and all parameters are estimated from the data itself. Data is provided as a by-product of the motion estimation procedure and the estimation of the probabilities is incorporated in the multiscale scheme. The method does not require the estimation of spatial or temporal derivatives which can introduce additional noise.

The remainder of this subsection is organized as follows. First, we express the block-matching motion estimation scheme in a probabilistic framework as a Maximum Likelihood estimator. Based on the objective criterion of the block-matching estimator assumptions are made about the type of the probability distribution of the motion-compensated intensity differences. Then, based on these assumptions, we derive the confidence measures in terms of the *a posteriori* probability of the motion vectors and present experimental results for image sequences for which the motion is known (appendix D).

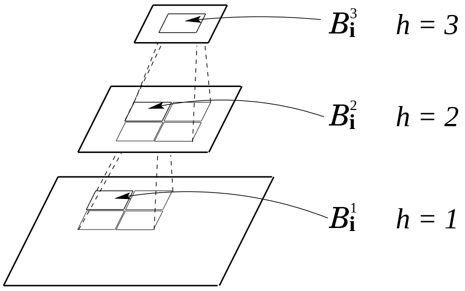


Figure 3.6: Hierarchical block-matching motion estimation: Block $B_{\mathbf{i}}$ and its ancestors $B_{\mathbf{i}}^h$ at level h

Block-Matching Motion Estimation in the Probabilistic Framework

Block-based motion estimators belong to a general class of estimators that utilize the **Intensity Conservation Principle**, that a pixel and its correspondence in a successive frame are expected to have the same intensity value. They attempt to overcome the *ill-posedness* of the correspondence problem by adopting a support region in the form of a block, and estimate a motion vector for the whole block. The motion vector is estimated as the one that minimizes an objective criterion which, typically, is either the Mean-Absolute Displaced Block Difference or the Mean-Square Displaced Block Difference. While the analysis that follows is based on the former objective criterion, it is almost straightforward to derive the confidence measures for the latter. Formally, at each level h the motion vector $\hat{\mathbf{v}}_{\mathbf{i}}^h$ is estimated for each block $B_{\mathbf{i}}^h$ (fig. 3.6) such that:

$$\hat{\mathbf{v}}_{\mathbf{i}}^h = \arg\min_{\mathbf{v}^h} D_{\mathbf{i}}^h(\mathbf{v}^h) \tag{3.3}$$

where,

$$D_{\mathbf{i}}^h(\mathbf{v}^h) = \sum_{\mathbf{j} \in B_{\mathbf{i}}^h} \left| I(\mathbf{j}) - I^-(\mathbf{j} - \mathbf{v}^h) \right| \tag{3.4}$$

With $\mathbf{i}$ we denote the pixel in the center of the block, and with $B_{\mathbf{i}}^h$ the set of the pixels in the block. With I and I^- we denote the image intensities in the current and previous frame respectively.

In what follows we will prove that the block-based motion estimator as defined by eq. 3.3 is equivalent to a Maximum Likelihood estimator under the assumption that the motion-compensated intensity differences follow independent Laplacian distributions. Under this assumption, the conditional probability of the intensities in the previous frame, given the motion vector $\mathbf{v}$ for the block $B_{\mathbf{i}}$ and the intensities in the current frame is given by:

$$P(I^-|I, \mathbf{v_i} = \mathbf{v}) = \prod_{\mathbf{j} \in B_\mathbf{i}} p(I^-(\mathbf{j} - \mathbf{v_i})|\mathbf{v_i} = \hat{\mathbf{v}}, I(\mathbf{j})) \tag{3.5}$$

$$= \prod_{\mathbf{j} \in B_\mathbf{i}} \frac{\lambda}{2} e^{-\lambda |I(\mathbf{j}) - I^+(\mathbf{j}+\mathbf{v})|} \tag{3.6}$$

$$= \left(\frac{\lambda}{2}\right)^{|B_\mathbf{i}|} e^{-\lambda D_\mathbf{i}(\mathbf{v})} \tag{3.7}$$

where the dependence on the level h of the multiscale scheme is omitted for notational simplicity. Since $\mathbf{v}$ is independent of λ it is straightforward that the maximization of the likelihood given by eq. 3.7 with respect to $\mathbf{v}$ is equivalent to the minimization described in eq. 3.3.

With the above derivations we have proven that the minimization performed by the block-based estimator (eq. 3.3) is equivalent to the Maximum Likelihood estimation. The parameter λ in eq. 3.6 is related to the deviation of the distribution. As such, it does not influence the location of the minimum but only the “width” of the distribution and therefore its value does not influence the motion estimation under the Maximum Likelihood criterion. On the other hand, the larger the deviation is, the lower is the relative importance of the differences in the objective criterion. The latter is depicted in fig. 3.7 where the likelihood ratio of two different candidate motion vectors is drawn as a function of the λ^{-1}. Each curve depicts the likelihood ratio for a certain value of the difference in the corresponding $D(\mathbf{v})$s (i.e. $D(\mathbf{v1}) - D(\mathbf{v2}) = d$, for different d). It is apparent, that the larger the deviation is, the closer the likelihood ratio is to one. Thus, the larger the deviation is, the less is our confidence in the candidate motion vector that generates the smaller of the $D(\mathbf{v})$. Let us note that the likelihood ratio can be useful as a confidence measure only when precisely two candidate vectors are available and was introduced here mainly in order to illustrate the importance of a good estimation of the parameter λ.

In general, the statistics of the motion-compensated intensity differences depend on the image sequence in question. In what follows, we make the hypothesis that they also depend on the local intensity variation. Our hypothesis is that λ is correlated to the amount of texture in the block and more specifically that it is inversely linear to the standard deviation of the intensity in the block in question (i.e. $\lambda_{B_\mathbf{i}} = \frac{\beta}{\sigma_{B_\mathbf{i}}}$). Although the linearity is not guaranteed, it is to be expected that the higher the degree of texture, the higher the variation in the observed motion-compensated intensity differences. This hypothesis has been tested in a number of image sequences and here we present results for two of them (**Y1** and **R1**) in which the degree of texture and the type and magnitude of motion vary significantly as explained in appendix D. In fig. 3.8 we present the inverse of an estimation of λ as a function of the within-block deviation of the intensity σ_B. In the same figure we present the lines fitted with the Least Squares criterion, on the one hand under our hypothesis and on the other hand under the usual assumption that λ is invariant to the local intensity deviation. It is clear that λ is highly correlated to the within-block intensity deviation and therefore our modeling follows more closely the true statistics of the motion-compensated intensity

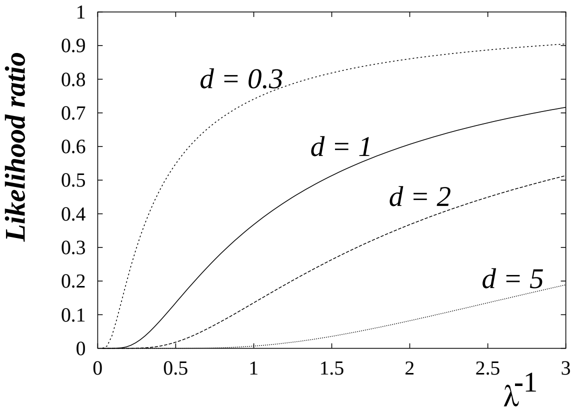


Figure 3.7: Likelihood ratio of two candidate motion vectors $\mathbf{v1}$ and $\mathbf{v2}$ as a function of the inverse of λ. Curves are drawn for different values of the differences in the corresponding $D(\mathbf{v})$ ($d = D(\mathbf{v1}) - D(\mathbf{v2})$)

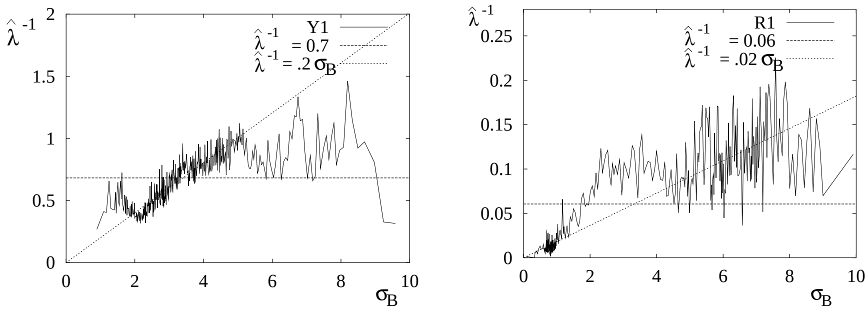
differences. In comparison, fig. 3.8 reveals that at areas with low intensity variation, the deviation of the underlying distribution is clearly underestimated. Finally, let us note the difference in scale between fig. 3.8(a) and fig. 3.8(b) and of the parameters of the fitted lines, which advocates the need for the re-estimation of the factor β for different image sequences.

Confidence measures

The block-matching motion estimator minimizes eq. 3.3 with a search scheme that evaluates different candidate motion vectors ($\mathbf{v}$) in terms of the objective criterion. In what follows, we will utilize the data provided by them in order to estimate the *a posteriori* probability of the best (and therefore chosen) candidate motion vector ($\hat{\mathbf{v}}_\mathbf{i}$). By using the theorem of Bayes and the total probabilities [88] we have that:

$$P(\mathbf{v_i} = \hat{\mathbf{v}}_\mathbf{i} | I^+, I) = \frac{P(I^+ | \mathbf{v_i} = \hat{\mathbf{v}}_\mathbf{i}, I) P(\mathbf{v_i} = \hat{\mathbf{v}}_\mathbf{i})}{\sum_{\mathbf{v}} P(I^+ | \mathbf{v_i} = \mathbf{v}, I) P(\mathbf{v_i} = \mathbf{v})} \tag{3.8}$$

where with $P(\mathbf{v_i} = \mathbf{v})$ we denote the *a priori* probability of a motion vector. With an appropriate modeling we can incorporate domain knowledge and/or utilize smoothness constraints. However, such issues which are not addressed in the classical block matching algorithm are not considered also in our analysis which assumes a uniform *a priori* distribution. Let us for notational simplicity denote with $g_\mathbf{i}$ the *a posteriori*



(a) **Y1** (b) **R1**

Figure 3.8: The inverse of the estimation of λ (i.e. $\hat{\lambda}^{-1}$) as a function of the within-block standard deviation of the intensity σ_B and the LS line fitting under the assumptions (i) that λ is constant and (ii) that it is inversely proportional to σ_B

probability. Then eq. 3.8 and eq. 3.7 imply that:

$$g_{\mathbf{i}} = \left(1 + \sum_{\mathbf{v} \neq \hat{\mathbf{v}}_{\mathbf{i}}} e^{-\lambda_{B_{\mathbf{i}}}(D_{\mathbf{i}}(\mathbf{v}) - D_{\mathbf{i}}(\hat{\mathbf{v}}_{\mathbf{i}}))}\right)^{-1} \tag{3.9}$$

$$= \left(1 + \sum_{\mathbf{v} \neq \hat{\mathbf{v}}_{\mathbf{i}}} e^{-\frac{\beta}{\sigma_{B_{\mathbf{i}}}}(D_{\mathbf{i}}(\mathbf{v}) - D_{\mathbf{i}}(\hat{\mathbf{v}}_{\mathbf{i}}))}\right)^{-1} \tag{3.10}$$

Note that the estimation of the *a posteriori* probability, as defined in eq. 3.10, depends on the search scheme that the particular block matching motion estimator employs. Loosely speaking, it gives a measure of how prominent the local minimum in $\hat{\mathbf{v}}_{\mathbf{i}}$ is, in terms of how much lower $D_{\mathbf{i}}(\mathbf{v}_{\mathbf{i}})$ is than the $D_{\mathbf{i}}(\mathbf{v})$ of the other candidate motion vectors.

As far as the parameter β is concerned, its value is assumed to be characteristic of the current frame, that is, its value is assumed to be the same for each block. It is estimated as:

$$\hat{\beta} = \mathcal{E}\left\{\frac{\sigma_{B_{\mathbf{i}}}}{D_{\mathbf{i}}(\mathbf{v}_{\mathbf{i}})}\right\} \approx \overline{\left(\frac{\sigma_{B_{\mathbf{i}}}}{D_{\mathbf{i}}(\hat{\mathbf{v}}_{\mathbf{i}})}\right)} \tag{3.11}$$

that is, as the mean value of the ratio of $D_{\mathbf{i}}(\hat{\mathbf{v}}_{\mathbf{i}})$ with the standard deviation of the intensity in block $B_{\mathbf{i}}$. Clearly, the mean is estimated over all blocks $B_{\mathbf{i}}$. Note that the estimation is with the Least Squares criterion but more robust estimates can be easily obtained. Since the images at each level are filtered and subsampled versions of the ones at the lower level, the statistics of the motion-compensated intensity differences differ from level to level and the parameter β is re-estimated at each level. Furthermore, the *a posteriori* probabilities provide a measure of confidence in the estimation at each level of the hierarchy. Clearly, there is a need to combine the evidence provided at each level in order to derive a measure that expresses the confidence in the multiscale motion estimation. Let us for notational simplicity denote with $g_{\mathbf{i}}^h$ the *a posteriori* probability $P(\mathbf{v_i}^h = \hat{\mathbf{v}}_{\mathbf{i}}^h | I^+, I)$ of the component of the motion vector estimated in level h. If the estimations at each level are considered independent, the probability that all of the $\hat{\mathbf{v}}_{\mathbf{i}}^h$ $(1 \leq h \leq H)$ are estimated correctly is equal to the product of $g_{\mathbf{i}}^h$. Experimentally however, we have found that the mean value of the $g_{\mathbf{i}}^h$ is a slightly more reliable measure. Formally, the confidence measure is defined as:

$$c_{\mathbf{i}} = \frac{P(\mathbf{v_i}^0 = \hat{\mathbf{v}}_{\mathbf{i}}^0 | I^+, I)}{H} \sum_{h=1}^{H} g_{\mathbf{i}}^h \tag{3.12}$$

Note, that we introduced the assumption that the value of λ is inversely linear to the local intensity variation only in the last step of the derivations of the *a posteriori* probability (i.e. eq. 3.8 to eq. 3.10). Therefore, our method can be easily adapted to another modeling of λ as long as an estimation scheme for the parameters of the model can be provided in place of eq. 3.11.

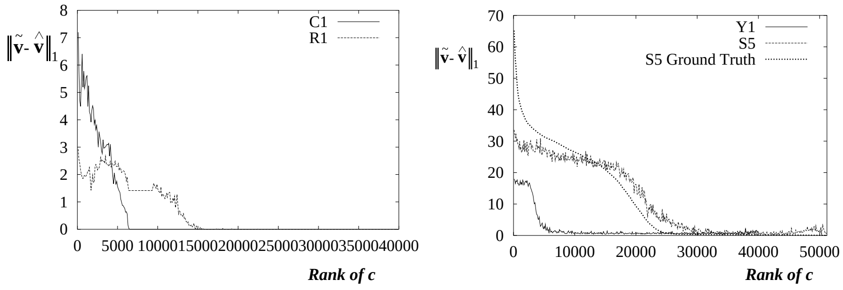
Computational Issues

The *a posteriori* probabilities at each level h are estimated from the Displaced Block Differences for each candidate motion vector. Except of the computationally inexpensive estimation of the deviation of the intensity within each block, the rest of the data that are used are a by-product of the search scheme. Therefore, from the data acquisition point of view we, practically, do not introduce any additional computational burden.

On the other hand, the estimation of the *a posteriori* probabilities is not possible until an estimation of the value of β is available. Practically, this means that the Displaced Block Differences need to be stored until β is estimated. The additional memory requirements depend naturally on the cardinality of the set of the motion vector candidates that the particular motion estimator employs. In the case that memory restrictions are too hard to allow such a scheme, an approximation can come by estimating the $g_{\mathbf{i}}^h$s using the value of β estimated for the previous frame and simultaneously estimating the value of β that will be used for the estimation of $g_{\mathbf{i}}^h$s for the next frame. Such an approach which assumes consistency of the statistics of the motion-compensated intensity differences in subsequent frames, introduces no additional computational costs or delays.

Experimental Results

In order to examine the validity of the derived measure we have conducted a number of experiments with synthetically generated data (appendix D). Here, we will present results for the image sequences **C1**, **R1**, **Y1** and **S5**. Each of the sequences exhibits different characteristics in terms of the amount of texture and the motion magnitude, as outlined in appendix D. The multiscale motion estimator was extended to half pixel accuracy and for the above-mentioned sequences 2, 1, 4 and 5 levels were used, respectively. Overlapping blocks were used and dense motion fields were derived.



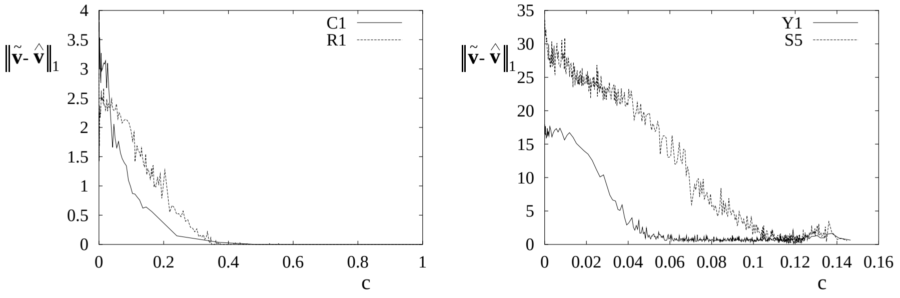
(a) **C1** and **R1**

(b) **S5** and **Y1**

Figure 3.9: Norm L_1 of the true motion estimation error (in pixels) as a function of the rank of the confidence measure c. For "S5 Ground Truth" the ranking is based on the true error instead of c

The original frames of the image sequences as well as the corresponding estimated motion fields are presented in appendix D. In fig. 3.9 we present the norm L_1 of the true estimation error (i.e. $||\hat{\mathbf{v}}_\mathbf{i} - \tilde{\mathbf{v}}_\mathbf{i}||_1$) as a function of the rank of $c_\mathbf{i}$ (100 pixels with consecutive ranks are used for each point in the plot). It is clear that the derived confidence measure is highly correlated to the true estimation error for all of the test sequences. As fig. 3.9 reveals, higher rank according to $c_\mathbf{i}$ implies, to a large extent, a small estimation error. In order to demonstrate the accuracy of the derived measure we present in fig. 3.9 the "ground truth" ranking for the image sequence **S5** ("S5 Ground Truth"). The curve S5, derived by ranking according to the proposed confidence measure, follows the "ground truth" curve quite closely. What is also clear, is the limitations that the measures that depend on rank information only have (e.g. [7]); the rise in error in "S5" starts at a much earlier point due to the larger degree of presence of occlusions. In fig. 3.10(a) and fig. 3.10(b) we present the norm L_1 of the error as a function of the confidence measure and in fig. 3.11. It is clear that the derived confidence measure adapts well to the presented sequences even though the type and magnitude of motion, the degree of texture and the degree of presence of occlusions differ significantly. For all of the sequences the only parameter (i.e. β) was estimated from the data by eq. 3.11.

Finally, in order to illustrate the localization accuracy, we present in fig. 3.11 and fig. 3.12 the norm L1 of the estimation error and the confidence measures as images for the sequences **R1** and **S5**, respectively. For illustration purposes the images are linearly stretched. It is clear that although the sequences **R1** and **S5** are very different, for both of them the structure of the image of confidence measures follows very closely the structure of the true error in the motion estimation. Note that the large areas in the **S5** sequence with low confidence are due to occlusions that originate from motions large in magnitude.



(a) **C1** and **R1**

(b) **S5** and **Y1**

Figure 3.10: Norm L_1 of the estimation error (in pixels) as a function of the confidence measure c

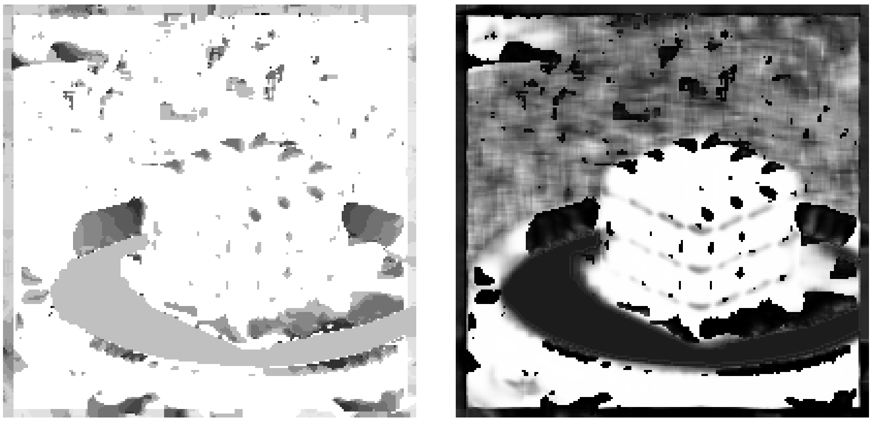
3.2.2 Clustering

Once the confidence measures are estimated, we incorporate them in a robust clustering technique in order to extract the set of motion hypotheses. Assuming that the number of motion hypotheses N is known, our goal is the estimation of $\Theta = \{\theta_n\} : n \in [1 \ldots N]$, where θ_n denotes the set of the affine parameters for motion hypotheses n (eq. 3.2). We do so by minimizing the following objective function with respect to Θ and L:

$$J(\Theta, L) = \sum_{s=1}^{M} |G_s| \rho\left(\sigma_c, r_s(\theta_{l_s})\right) \quad \text{, where} \quad r_s(\theta_{l_s}) = \sqrt{\frac{\sum_{\mathbf{i} \in G_s} c_{\mathbf{i}} r_{\mathbf{i}}^2(\theta_{l_{\mathbf{i}}}, l_{\mathbf{i}})}{|G_s|}} \tag{3.13}$$

where l_s is the cluster index (label) of the intensity segment s and G_s the set of pixels that belong to segment s. With $r_{\mathbf{i}}$, we denote the motion residual for pixel $\mathbf{i}$, which is the L_2 norm of the difference vector of the estimated motion vector $\hat{\mathbf{v}}_{\mathbf{i}}$ and the model-generated motion vector $\tilde{\mathbf{v}}_{\mathbf{i}}(\theta_{\mathbf{l_i}})$. The quantity $r_s(\theta_{l_s})$, which is defined in terms of the motion residuals of the pixels belonging to segment s can be loosely interpreted as a kind of a segment-based motion residual. Finally, $\rho()$ is a function the choice of which determines the shape of the objective function in terms of the relative influence of residuals of different magnitude. Two classical examples are the quadratic operator, which assigns a cost which is the square of the magnitude of the residual, and the Tukey bi-weight operator, which assigns a constant cost to residuals that are larger than a certain threshold.

In the case that (a) the confidence measures are ignored (i.e. $c_{\mathbf{i}} = 1$), (b) the function ρ is the quadratic function and (c) each intensity (color) segment consists of a single pixel, the proposed measure is the well-known L_2 norm used in the original C-Means algorithm [33]. By using the confidence measures $c_{\mathbf{i}}$ we limit the influence of the erroneous motion vectors, that is, the influence of the second type of residuals (p. 28). The function $\rho()$ belongs to the family of M-estimators [63], which are widely



(a) Linearly stretched error norm L_1. High intensity values indicate low true error

(b) Linearly stretched confidence values. High intensity values indicate high confidence values

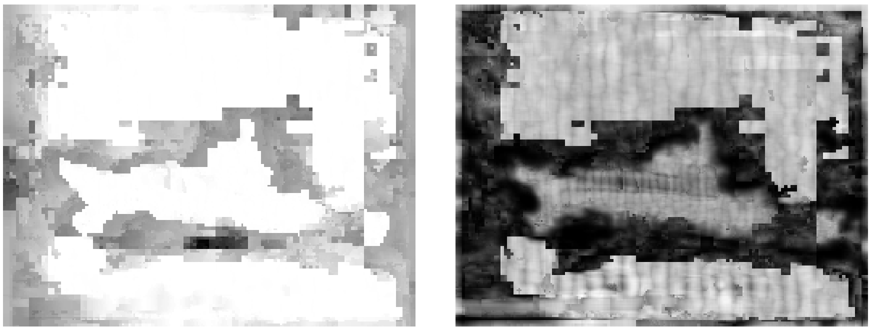
Figure 3.11: True error in motion estimation and confidence measure for image sequence **R1**

used in robust regression, clustering and reconstruction schemes (e.g. [63] [15]). Such functions are more robust in the presence of outliers, that is, of measures that do not conform to the assumed model. We have used the **Geman-McClure** function (eq. 3.14) a plot of which is depicted in fig. 3.13 for different values of the scale parameter σ_c.

$$\rho(\sigma_c, x) = \frac{x^2}{\sigma_c^2 + x^2} \tag{3.14}$$

As far as the role of scale parameter is concerned let us note that for $\sigma_c \rightarrow \infty$ the $\rho()$ tends to be equivalent to the quadratic operator. Practically, since the magnitude of the estimated motion vectors is at most in the order of tens of pixels, $\rho()$ is essentially equivalent to the quadratic operator for a value of σ_c equal to the larger motion residual (i.e. in the order of ten). At the other extreme, a very small value of σ_c penalizes practically equally every residual, independently of its magnitude.

In the context of our method, the role of the M-estimator $\rho()$ [63] is to robustify the estimation step of the modified C-Means algorithm, that is, the optimization with respect to the motion hypotheses Θ, by reducing the influence of (a) wrongly classified intensity segments and (b) segments which contain large residuals that their influence was not reduced enough by the confidence measures. Finally, the proposed method imposes a common fate on pixels who belong to the same intensity segment. As a result, the computational complexity is drastically reduced, as will become apparent when the optimization procedure is described.



(a) Linearly stretched error norm L_1. High intensity values indicate low true error

(b) Linearly stretched confidence values. High intensity values indicate high confidence values

Figure 3.12: True error in motion estimation and confidence for image sequence **S5**

Optimization

In order to minimize $J(\Theta, L)$, we follow a procedure which iterates between an estimation and a classification phase. In the classification phase a minimization with respect to L takes place, where the motion hypotheses that are used were estimated in the previous iteration. Since no interdependencies exist in the label field L, that is, the label l_s of a segment s does not influence the label of any other segment s', the labeling is performed independently for each segment. Trivially, each segment is assigned the object label n that minimizes the cost under the motion hypothesis θ_n. Formally,

$$l_s = \arg\min_n \rho(\sigma_c, r_s(\theta_n)) \tag{3.15}$$

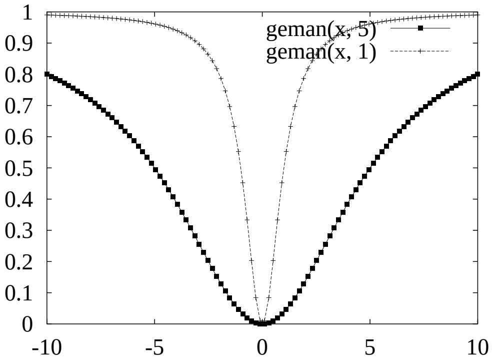


Figure 3.13: Function Geman for $\sigma_c = 5$ and $\sigma_c = 1$

In the estimation phase, a minimization with respect to Θ takes place. The solution is given by differentiating $J(\Theta, L)$ with respect to each set of parameters θ_n. After some derivations:

$$\begin{aligned} \nabla_{\theta_n} J(\Theta, L) = 0 &\Longrightarrow \sum_{\{s:l_s=n\}} \left(w_s(r_s(\theta_n)) \sum_{\mathbf{i}\in G_s} c_{\mathbf{i}} \nabla_{\theta_n} r_{\mathbf{i}}(\theta_n)^2 \right) = \mathbf{0} \\ &\Longleftrightarrow \sum_{\{\mathbf{i}:l_{\mathbf{i}}=n\}} w_s(r_s(\theta_n)) \nabla_{\theta_n} r_{\mathbf{i}}(\theta_n)^2 = \mathbf{0} \end{aligned} \tag{3.16}$$

where $w_s(x) = \frac{1}{x}\frac{d\rho(x)}{dx}$. eq. 3.16 is non-linear with respect to θ_n and a closed formed solution cannot be obtained. A usual way of solving eq. 3.16 is the Iterative Re-weighted Least Squares (IRLS) method [46], which iterates between a calculation of w_ss given θ_n and an estimation of θ_n given the w_ss. As its name implies, the latter is equivalent to the least-squares solution, if each equation is weighted with the corresponding weight w_s. That means that the system which needs to solved at each iteration of the IRLS method is linear with respect to the parameters θ_n.

Since the optimization procedure is essentially a gradient descent technique, convergence is guaranteed only to a local minimum which depends on the initializations. Furthermore, the lower the value of the scale parameter σ_c, the smaller the influence of segments that do not conform with the already estimated motion parameters to their re-estimation; therefore, the higher the danger of getting trapped in a local minimum. In order to deal with the latter, we adopt a scheme in which σ_c is gradually reduced. The scheme corresponds to least-squares estimations in the first iterations and gradually reduces the influence of larger residuals.

Even with the use of an M-estimator, the dependency of the iterative clustering scheme on the initial values of the motion parameters, or equivalently their dependency on the initial labeling remains. In order to deal with it, we apply the clustering scheme for a number of epochs with different initializations and choose the best in terms of the objective function $J(\Theta, L)$. The value of the scale parameter σ_c that is used for such a validation is estimated as the normalized weighted standard deviation of the motion residuals at the iteration which yields the smaller normalized weighted standard deviation of the motion residuals. For the initialization of the parameters at each epoch of the clustering scheme we use the following procedure. A random variable is associated to each of the affine parameters as depicted in Table 3.1. Each random variable corresponds to the solution with respect to the parameter in question, under the constraint that all the other parameters are zero. The initialization is then obtained by uniform sampling in the hyper-parallelepiped which is defined by the second order moments of the random variables. For example, the parameter $\theta(1)$ for the cluster n is initialized by uniform sampling in the range $[-2\sigma_{u_x} \ldots 2\sigma_{u_x}]$, where σ_{u_x} is the standard deviation of the horizontal component of the motion vectors.

Finally, as far as the computational complexity is concerned, we should note that the formulation of the objective function allows a computational complexity of $O(M)$ at each iteration, where M is the number of intensity segments. That is, the complexity of the re-estimation with respect to Θ is $O(M)$ and of labeling a segment s is $O(1)$. In comparison, the corresponding pixel-based clustering methods exhibit a computational

Affine Parameter	$\theta(1)$	$\theta(2)$	$\theta(3)$	$\theta(4)$	$\theta(5)$	$\theta(6)$
Random Variable	$\frac{\mathbf{v}_x}{\mathbf{i}_x}$	$\frac{\mathbf{v}_x}{\mathbf{i}_y}$	$\mathbf{v}_x$	$\frac{\mathbf{v}_y}{\mathbf{i}_x}$	$\frac{\mathbf{v}_y}{\mathbf{i}_y}$	$\mathbf{v}_y$

Table 3.1: Affine parameters and the associated Random Variables. $\mathbf{v}_x$ and $\mathbf{v}_y$ denote the horizontal and the vertical motion component, respectively

complexity in the order of the number of pixels.

In order to achieve the mentioned computational complexity we need to estimate intermediate matrices that allow (a) the computation of the segment motion residual $r_s(\theta)$ in $O(1)$ and (b) the construction of the linear systems that result from eq. 3.16 in $O(M)$. In order to do so let us define $a_x = [\theta(1)\ \theta(2)\ \theta(3)\ -1]^T$, $u_x(\mathbf{i}) = [\mathbf{i}_x\ \mathbf{i}_y\ 1\ \mathbf{v}_x]$ and similarly a_y and u_y. Then

$$r_s(\theta)^2 = a_x^T \left(\sum_{\mathbf{i} \in G_s} c_{\mathbf{i}} u_x(\mathbf{i}) u_x(\mathbf{i})^T \right) a_x + a_y^T \left(\sum_{\mathbf{i} \in G_s} c_{\mathbf{i}} u_y(\mathbf{i}) u_y(\mathbf{i})^T \right) a_y \tag{3.17}$$

and the solution with respect to $[\theta_n(1)\ \theta_n(2)\ \theta_n(3)]$, as derived from eq. 3.16, is given by the 3×3 system that is defined from the first three equations of

$$\left(\sum_{s: l_s = n} w_s \sum_{\mathbf{i} \in G_s} c_{\mathbf{i}} u_x(\mathbf{i}) u_x(\mathbf{i})^T \right) a_x = 0 \tag{3.18}$$

Since the matrix $\sum_{\mathbf{i} \in G_s} c_{\mathbf{i}} u_x(\mathbf{i}) u_x(\mathbf{i})^T$ depends neither on the affine parameters nor on the labeling it can be estimated once for each segment. Then, the calculation of r_s reduces to a vector-matrix multiplication and the re-estimation of θ to an addition of M matrixes and solving two 3×3 linear systems.

Experimental results

The motion extraction method has been tested in a large number of image sequences, both synthetic and real, and results are presented for the proposed algorithm as well as for methods derived from some combinations of its degenerate cases listed below.

(a) Confidence measures are not used ($c_{\mathbf{i}} = 1$)

(b) Each segment consists of a single pixel ($|G_s| = 1$)

(c) The quadratic function is used instead of the M-estimator ($\sigma_c \to \infty$). This is equivalent to the Least Squares solution.

In what follows, we will present results for the proposed method (i.e. E) and its special cases (i.e. B-D) as these are depicted in Table 3.2.

For the synthetic sequence **S5** where the ground truth is known, we present results in terms of accuracy in the estimated parameters. The results are summarized

Method name	Constraints	Description
A	**(a)** & **(b)** & **(c)**	C-Means [33]
B	**(b)**&**(c)**	Confidence based C-Means
C	**(a)** & **(c)**	Segment-based C-Means clustering [3]
D	**(c)**	Confidence and segment-based C-Means
E		Proposed method

Table 3.2: Proposed clustering method and its special cases

Method	$\theta(1)$	$\theta(2)$	$\theta(3)$	$\theta(4)$	$\theta(5)$	$\theta(6)$	σ_{L_2}
Ground Truth	0	0.03	-2	0.1	0	-13	
A	0.0048	-0.0352	-0.76	0.056	-0.022	-7.13	10.05
B	0.0042	-0.0203	-1.01	0.055	0.084	-12.62	≈ 0
C	0.0116	-0.0560	0.87	0.029	-0.052	-5.73	≈ 0
D	0.0149	-0.0202	-0.39	0.054	0.094	-13.21	6.88
E	0.0007	0.0284	-1.95	0.10	0.002	-13.05	≈ 0

Table 3.3: Estimated motion parameters for **S5** image sequence

in Table 3.3 where the true and the estimated parameters are presented for the “object” (fig. D.4(a)). Equivalent, but slightly better estimates are obtained for the “background”. The main characteristic in this sequence is the large magnitude of motions which generate large occlusions and residuals large in magnitude (fig. 3.12).

It is clear from Table 3.3 that the incorporation of confidence measures in the clustering procedure significantly improves the accuracy of the estimation and that, as expected, the contribution is more significant for the segment-based methods. The last column of Table 3.3 depicts the variance of the weighted mean L_2 norm of the motion residual estimated over the different epochs. That serves as a measure of the dependence of the corresponding method on the initial values of the parameters.

Finally, let us note that the computational advantage of the segment-based methods is not only due to the lower complexity within each iteration (a factor of 84), but also due to the fact that convergence is achieved in fewer iterations. This is because that the solution space for the label field is drastically reduced in comparison to that in pixel-based methods. While for the first two of the methods presented in Table 3.3 convergence requires around 35 and 20 iterations respectively, the corresponding segment-based methods converge in around 12 and 9 iterations respectively. Even when the additional operations due to vector-matrix multiplications are taken into account, this leads to a total reduction in the computational complexity of a factor around 50.

For real image sequences where the ground truth is not known, the results are evaluated according to the labeling as well as according to the stability. Here, we will present results for the image sequences “train”, “sunflower garden” and “coast guard”.

The image sequence “train” is characterized by translational motions of large magnitude that generate large occlusions. In fig. 3.14 the 10th frame and the confidence measures on the corresponding estimated motion field are presented. A half-pixel ac-

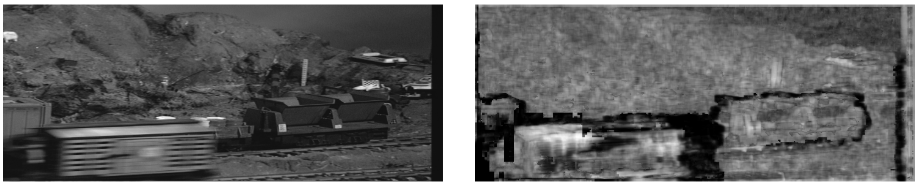


Figure 3.14: 10th frame and the confidence measures of the corresponding estimated motion field for “train” sequence

curacy motion estimator was used. The motion vectors that, due to occlusion phenomena, are unreliable are correctly identified. In fig. 3.15 we present the label fields at different stages of the optimization procedure for a typical good epoch of the proposed algorithm. It is difficult to evaluate the accuracy of the estimated motion parameter since there is no ground truth for this sequence. On the other hand the proposed method identifies quite reliably the region of support for the three dominant motion patterns. As will become apparent when the labeling stage is described, the extracted motion hypotheses are accurate enough to permit reliable labeling.

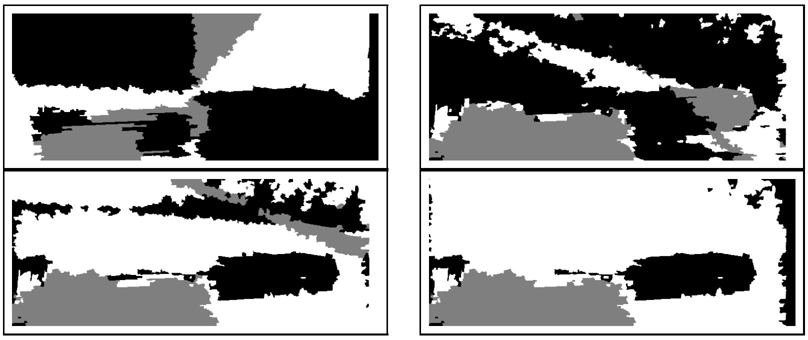
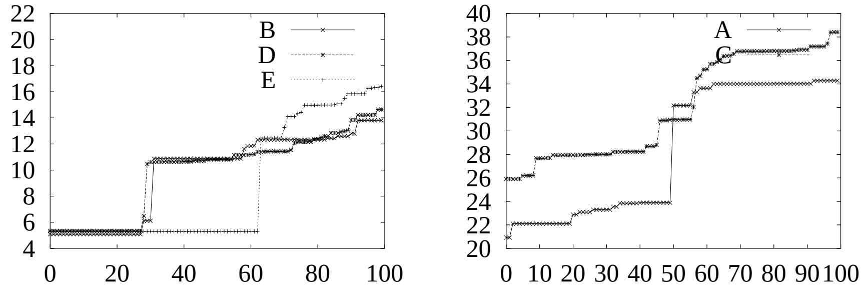


Figure 3.15: Evolution of the label field for “train” sequence. (Iteration 1, 3, 9 and 121)

In comparison to its special cases the proposed algorithm proves considerably more stable. The comparative results are summarized in fig. 3.16, where the weighted L_2 norm of the motion residual (denoted with wL_2) is presented for each of the five methods that are outlined in Table 3.3. For illustration purposes the epochs are sorted according to wL_2. What is relevant in fig. 3.16 is the number of epochs for which a low error norm is achieved. The results clearly illustrate that the methods that utilize the proposed confidence measure outperform the corresponding methods that do not by far. While the methods B and D converge to the solution of fig. 3.15 around 28% of the epochs, the corresponding methods A and C achieve a similar solution for 20% and 10% of the epochs, respectively. Furthermore, for the methods that do not utilize the confidence measures the objective function seems not so well behaved. There exist many local minima which are very similar in terms of the objective function but far in terms of the motion parameters. In order to illustrate the latter, we present in fig. 3.17

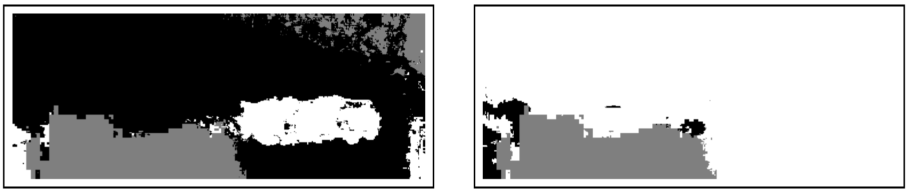


(a) Methods B, D and E

(b) Methods A and C ($c_i = 1$)

Figure 3.16: Weighted L_2 norm (wL_2) of the motion residual estimated at convergence at subsequent epochs. Large deviations from the minimum of each method indicate larger dependence of the method on the initialization of the motion parameters. For visibility purposes the epochs are sorted according to wL_2

the label fields for two different epochs where the corresponding mean motion residuals at convergence were 22.10 and 22.88 respectively. Although the mean motion residuals at convergence are very similar, the corresponding label fields (and therefore the corresponding motion hypotheses) differ greatly. In comparison, a similar degradation in terms of the objective criterion for the methods B and D consistently results in a label field that assigns the label of the train in the foreground to the visible part of the sky (upper right corner of the image). Finally, the proposed algorithm converges exactly to the solution of fig. 3.15 for 63% of the epochs, more than twice as often as the second best of the examined methods.



(a) Label field at convergence $L_2 = 22.10$

(b) Label field at convergence $L_2 = 22.88$

Figure 3.17: Clustering results for two epochs of C-Means with very similar L_2's at convergence

As far as the computational complexity is concerned, the results are summarized in fig. 3.18. The initial intensity segmentation results in 2338 intensity segments, that is, the computational complexity at each iteration is reduced by a factor of 88 for the segment-based methods with respect to the pixel-based methods. When the quadratic function is used instead of the M-estimator the computational complexity is reduced by an additional factor of around 5 due to faster convergence in terms of iterations. However, once the M-estimator is used, the schedule by which σ_c is gradually lowered requires a fixed number of iterations which is on average three times larger than the number of iterations required by the pixel-based methods. The total reduction in computational complexity of the proposed method in comparison to pixel-based methods (taking into account the additional cost due to vector-matrix multiplications) is of a factor of roughly 10.

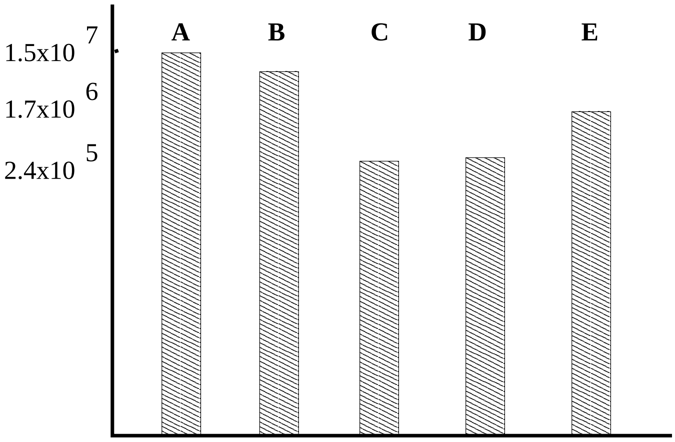
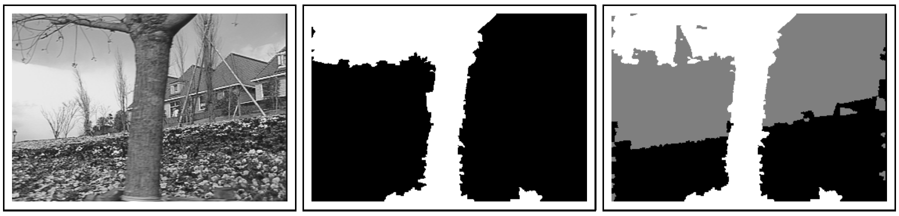


Figure 3.18: Computational complexity of the methods (logarithmic scale)

Finally, we present results for the image sequence “sunflower garden” (fig. 3.19) where the apparent motion is generated due to camera motion. The different motion patterns are due to the relative difference in the distances of the objects from the camera. For this sequence we have conducted experiments for $N = 2$ and $N = 3$. In the former case, the behavior of the different methods is considerably more stable than in the case of the “train” sequence, and these converge almost consistently to solutions almost identical to the one presented in fig. 3.19(b). However, when $N = 3$, the proposed algorithm is the only one that converges to the solution presented in fig. 3.19(c); the rest of the methods fail to separate successfully the background.

3.3 Labeling Using Markov Random Field Theory

At the second stage of our method (fig. 3.1), the segments are labeled anew given the set of motion parameters Θ as estimated in the motion hypotheses extraction phase. In comparison to the internal labeling in the clustering phase (section 3.2.2) there are two main differences. The first difference is that the labeling does not depend anymore on the estimated motion field. While, for the purpose of the motion hypotheses extraction the inaccuracies of the motion field can be tolerated by using confidence measures and an M-estimator, this is not the case for the purpose of labeling. If, for any reason, the



(a) Frame 10 (b) $N = 2$ (c) $N = 3$

Figure 3.19: Frame 10 and the corresponding label fields at convergence of the clustering for the “sunflower garden” sequence

motion vectors within a segment are wrongly estimated, the segment is most likely to be misclassified. Especially in occluded areas, misclassifications are to be expected. Instead, we rely on the motion-compensated intensity differences which do not introduce the artifacts that a feature estimation procedure introduces. The second difference is that we introduce spatial constraints by modeling appropriately the label field; such constraints were absent in the labeling of section 3.2.2.

More specifically, we aim at the maximization of the *a posteriori* conditional probability of the label field, given the estimate of the motion hypotheses ($\hat{\Theta}$) and the image intensities in the previous, current and next frame (denoted with I^-, I and I^+ respectively). Using Bayes rule

$$P\left(L|I,\hat{\Theta},I^-,I^+\right) \propto P\left(I|L,\hat{\Theta},I^-,I^+\right) P\left(L|\hat{\Theta}\right) \tag{3.19}$$

The maximization of eq. 3.19 is equivalent to the minimization of the negative of its logarithm, that is, the sum of the logarithms of each of the two terms on the right-hand side. The first term on the right-hand side of eq. 3.19 expresses how well the current motion and label field conform to the image intensities. As in the motion estimation phase (eq. 3.6) we assume that the motion-compensated intensity differences follow a Laplacian distribution. Then:

$$P\left(I|L,\hat{\Theta},I^-,I^+\right) = \prod_{s=1}^{M}\prod_{\mathbf{i}\in G_s} \lambda_{l_s} e^{-\lambda_{l_s}\min\{|f_{\mathbf{i}}^-(\hat{\theta}_{l_s})|,|f_{\mathbf{i}}^+(\hat{\theta}_{l_s})|\}} \tag{3.20}$$

where with $f_{\mathbf{i}}^-(\theta)$ and $f_{\mathbf{i}}^+(\theta)$ respectively, we denote the backward and forward motion-compensated intensity difference under the motion hypothesis θ at pixel $\mathbf{i}$. With λ_{l_s} we denote the parameter λ of the Laplacian that models the motion-compensated intensity differences for the object with label l_s.

Note that we use a three-frame approach where at each pixel either the backward or the forward motion-compensated intensity difference is considered valid. The underlying assumption is that each pixel has a correspondence either in the previous or in the next frame. By doing so, we deal in a simple and efficient way with occlusions. A similar approach is presented by Dubois and Konrad[32]. In their work the direction in which the motion estimation is constrained is determined by an occlusion field.

The second term on the right-hand side of eq. 3.19 models the probability of the label field. We model it as a Gibbs distribution whose energy term $E_c(L)$ is the sum of spatial clique potentials $V_c\left(s, s'\right)$ which are defined over pair-site (pair-segment) cliques as follows:

$$V_c\left(s, s'\right) = \begin{cases} -z_c b\left(s, s'\right) & \text{if } l_s = l_{s'} \\ z_c b\left(s, s'\right) & \text{if } l_s \neq l_{s'} \end{cases} \tag{3.21}$$

where the segments s and s' are neighbors in the neighborhood system N_s defined on the region adjacency graph. The term z_c is a constant that controls the weight of the spatial constraints relative to the constraints that the intensity preservation principle imposes. The term $b\left(s, s'\right)$ denotes the length of the common border between s and s', which is estimated as the number of pairs of pixels $(\mathbf{i}, \mathbf{i}')$ which are neighbors in the image grid and belong to the borders of s and s', respectively. From a global point of view, an optimization with respect to the spatial energy term $E_c(L)$ tends to minimize the total border length between neighboring objects.

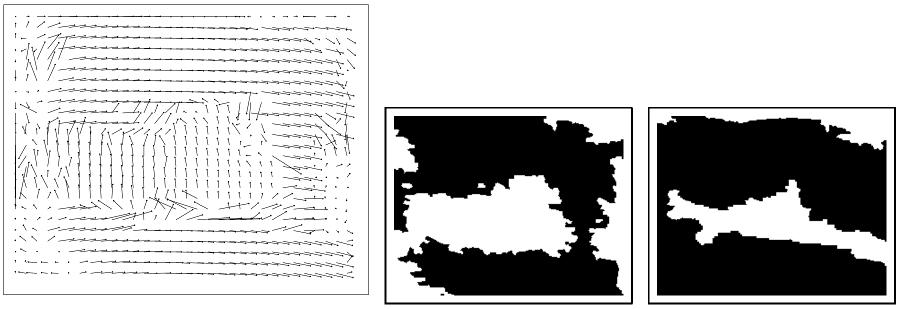
The parameter z_c controls the relative weights of the spatial constraints in the estimation of the label field, with respect to the data energy term E_d (E_d is equal to the negative logarithm of $P\left(I|L, \hat{\Theta}, I^-, I^+\right)$). As far as the spatial energy term is concerned, we note that at segment level the spatial energy term is proportional to the perimeter of the segment s, while the data energy term is proportional to the number of pixels of s. In general, this implies that the larger a segment, the larger the ratio between the relative contribution of the local data energy term with respect to the local spatial energy term. Thus, for larger segments emphasis is given to the evidence that the segments themselves provide about their temporal behavior, while for smaller segments the emphasis is given to the evidence provided by the label field in their neighborhood. For the manual setting of the value of z_c a rule of thumb can be provided by an analysis of the ratio between the size and the perimeter of the segments. For a value of z_c around 1.5 a rough normalization between the different energy terms is achieved for segments of around 100 pixels in size. In our experiments we used values of z_c in the range between 2 and 12.

For the optimization procedure we employ a deterministic relaxation algorithm known as Iterative Conditional Modes (ICM) [11]. In an iterative way, ICM maximizes at each iteration the conditional probability of a label each site (intensity segment) given the labels at all other sites. Since the Markov Random Field modeling introduces only local dependencies, this conditional probability can be expressed in terms of the local energy terms. More specifically:

$$\begin{aligned} &-\ln P(l_s = n | I, I^-, I^+, \hat{\theta}_n, \{l_{s'} : s' \in N_s\}) = \\ &\qquad \lambda_n \sum_{\mathbf{i} \in G_s} |f_{\mathbf{i}}\left(\hat{\theta}_{l_s}\right)| - |G_s| \ln \frac{\lambda_n}{2} + \sum_{s' \in N_s} V_c\left(s, s'\right) \end{aligned} \tag{3.22}$$

where $f_{\mathbf{i}}\left(\hat{\theta}_{l_s}\right) = \min\left\{|f_{\mathbf{i}}^-\left(\hat{\theta}_{l_s}\right)|, |f_{\mathbf{i}}^+\left(\hat{\theta}_{l_s}\right)|\right\}$. At each iteration only the term $V_c\left(s, s'\right)$ is affected, which implies that the computational complexity at each iteration is $O(\sum_{s=1}^{M} |N_s|)$, that is, it is proportional to the number of segment cliques. An initialization of the label field is provided by the labeling of each segment according to the Maximum Likelihood principle (i.e. $P(I, \hat{L}^{-}|L)$). After some derivations a ML maximization is equivalent to a minimization of all but the last term of eq. 3.22. Since there are no interdependencies the ML labeling can be performed independently per segment.

In order to investigate the validity of the proposed algorithm we have conducted a number of experiments with both synthetic and "real" image sequences. We will present here results for the synthetic image sequence "**S5**" and the real image sequences "train", "sunflower garden" and "coast guard".



(a) Estimated motion field

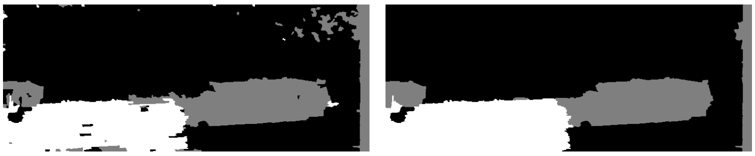
(b) Label field at convergence of the clustering

(c) Label field at convergence of the ICM

Figure 3.20: "S5" image sequence: Labeling based on motion field versus labeling based on motion-compensated intensity differences

In fig. 3.20 we present experimental validation of our choice in which a labeling based on the motion-compensated intensity differences was preferred over a classification of the estimated motion vectors. The estimated motion field for the synthetic image sequence "**S5**" (fig. 3.20(a)) has too many errors to allow for a reliable classification. Although the parameters for both the object and the background are estimated with high accuracy (Table 3.3) misclassifications occur near motion discontinuities, where the motion field is mainly corrupted. Note that the misclassifications are not only due to occlusions. The motion field was estimated in a backward manner (i.e. between the frame in question and the preceding frame) so that in principle the motion for the upper part of the object and the right margin of the image could be estimated accurately. Still, due to the limitations of the motion estimator, errors are present in the motion field, even at these areas, which lead to misclassifications. In comparison, the labeling based on the motion-compensated intensity differences and the Markov Random Field modeling labels successfully the areas near motion discontinuities (3.20(c)). Due to the bi-directional way in which the motion hypotheses are validated both occluded and disclosured areas are successfully labeled. The only areas in which there are differences with the true object mask (fig. D.4(b)) are areas that have correspondences in neither the previous nor in the next frame. For such areas it is very questionable if a correct classification based on a 3-frame analysis can be achieved.

The image sequence "train" is also characterized by motions large in magnitude and therefore the same analysis as the one for the "S5" sequence can be applied. In fig. 3.21(a) we illustrate the labeling based on the Maximum Likelihood principle that is used as an initialization and in fig. 3.21(b) the label field obtained at convergence by the Iterative Conditional Modes (ICM) algorithm. It is clear that by modeling the label field as a Markov Random Field we are able to obtain "clean" label fields without isolated segments. The localization accuracy of the method is illustrated in fig. 3.22, where the edges of the label field are superimposed on the original frame. The result justifies our choice of a conservative initial intensity segmentation; the borders of the objects are well preserved. The intensity constraints that are incorporated by the use of the initial segmentation allow us to obtain a label field where the edges are very well localized.



(a) Maximum Likelihood (b) MRF modeling

Figure 3.21: "train" image sequence: Labeling without and with spatial constraints

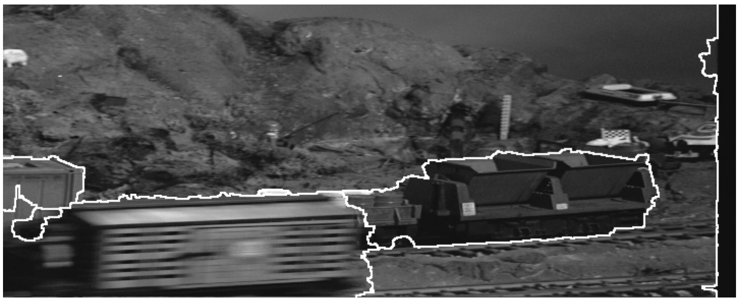
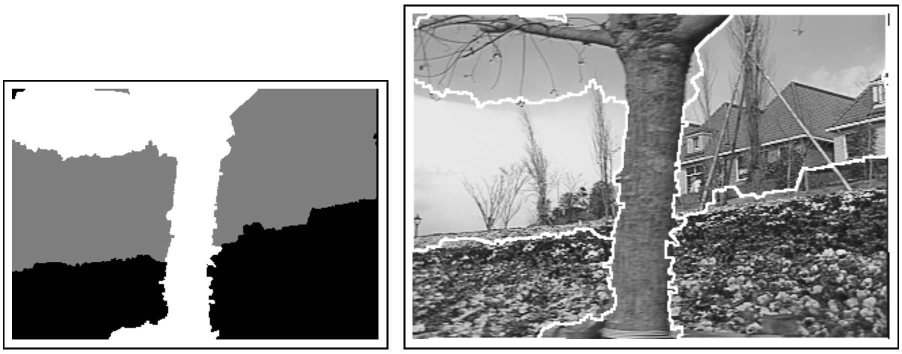


Figure 3.22: Object edges derived from the label field of fig. 3.21(b) superimposed on the original frame

Similar results were obtained for a number of image sequences. In fig. 3.23 we present the label field obtained at convergence for the "sunflower garden" image sequence for $N = 2$ and the edges of the label field superimposed on the original frame for $N = 3$. The trunk of the tree as well as its right branch are well localized but the thinner branches are not preserved by the initial intensity segmentation. As a consequence, part of the sky is labeled with the same label as the tree, since for the sky there is not enough texture to support the background motion hypothesis.



(a) Label field for $N = 3$

(b) Superimposed edges for $N = 3$

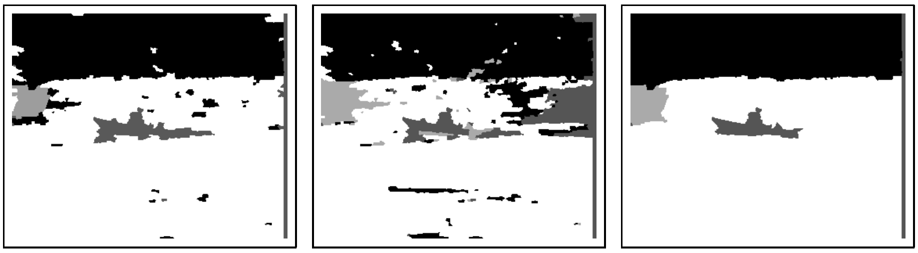
Figure 3.23: “sunflower garden” image sequence. Labeling and superimposed edges

Finally, we present results for the 10th frame of the “coast guard” sequence for which the difficulties arise due to the small magnitude of the different motion patterns and the complicated pattern of the motion of the water. The motion of the water, the middle boat and the shore can be distinguished only at sub-pixel level. The proposed motion hypotheses extraction method is able to successfully extract the four most dominant motion patterns as shown in fig. 3.24(a), where the label field obtained at convergence of the proposed clustering algorithm is depicted. Based on the extracted motion hypotheses, the initialization and the label field obtained at the convergence of the optimization procedure of section 3.3 are presented in fig. 3.24(b) and fig. 3.24(c) respectively. The localization accuracy for that sequence is illustrated in fig. 3.25. Both of the ships are separated well from the water; the bow waves that are created in front part of the ships also appear to move with them and therefore are not labeled as part of the water. Furthermore, the shore is also well separated from the water even thought the motion of the shore and the water differ only around half a pixel per frame.

3.4 Conclusions

In this chapter we presented a method for labeling image sequences based on the kinematic characteristics of the depicted objects. The proposed methodology approaches the problem in a two-stage way; the extraction of the kinematic characteristics is separated from the labeling itself.

At the first stage, motion hypotheses are extracted by cluster analysis of a motion field. Confidence measures developed specifically for the motion estimator that was used and robust regression techniques were utilized in a clustering scheme. The proposed confidence measure are derived by expressing the block-based motion estimation scheme in the probabilistic framework and are expressed in terms of the *a posteriori*



(a) Label field obtained at convergence of the clustering

(b) Maximum Likelihood labeling

(c) Label field obtained at convergence of ICM

Figure 3.24: Label fields for “coast guard” sequence

probability of the corresponding estimated motion vectors. We have experimentally supported our hypothesis that the deviation of the motion-compensated intensity differences is linear with the local intensity deviation. The method that we have proposed, practically introduces no additional computational burden since all of the data is provided as a by product of the search scheme of the block-matching motion estimator. We have illustrated the advantages of using confidence measures and robust regression techniques in terms of accuracy of the estimated motion parameters and in terms of stability. We have additionally constrained the clustering problem by utilizing an initial intensity segmentation scheme to form initial groups of points and reduced the computational complexity drastically. The efficiency of the proposed scheme has been illustrated for a number of image sequences in our experimental results.

At the second stage, the motion hypotheses are used to label the segments that result from the initial intensity segmentation. We have illustrated that by adopting a conservative approach it is possible to obtain in most cases an initial intensity segmentation that preserves the edges of the objects and we commented on the limitations of the method that we have chosen. In contrast to the majority of similar methods in the literature our method uses the motion-compensated intensity differences as evidence for the labeling. We have clearly illustrated the advantages of such an approach in the presence of motions large in magnitude. We have posed the labeling problem as an optimization problem in terms of the a posteriori probability of the label field, given the motion hypotheses. A three-frame approach was adopted and its contribution in dealing with occlusions was clearly illustrated. We have incorporated spatial constraints by modeling the label field as a Markov Random Field. We have illustrated the efficiency of the proposed method in a number of image sequences.

As far as future work in this direction is concerned, we would propose the use of a more efficient optimization scheme in the clustering so that multiple initializations are not necessary. The formulation can be easily adapted so that stochastic optimization is applied. A similar argument holds for the labeling procedure. Furthermore, the need of tuning the parameter z_c, the only parameter except of the number of objects, might be possible to be eliminated. The label field provided by the Maximum Likelihood

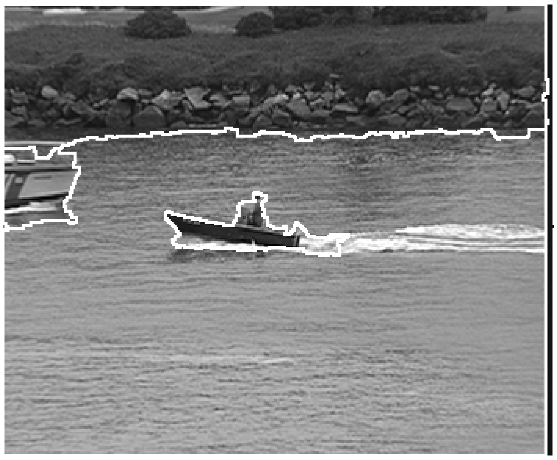


Figure 3.25: “coast guard” image sequence: Edges of the label superimposed on the original frame

classification is usually quite close to the desired one. The use of cross validation techniques might be sufficient to provide a robust solution to this issue.

Chapter 4

Joint Motion Estimation and Segmentation

In the previous chapter we presented a motion-based segmentation method in which the extraction of the motion hypotheses and the estimation of the label field are performed in separate stages. In the first stage a set of motion hypotheses were extracted by clustering an independently estimated motion field. In the second stage, segments that were extracted in an intensity segmentation phase were labeled according to the motion hypothesis to which they conformed. The method gave very good results for a number of image sequences but does not exploit the temporal redundancy either in the motion or in the label field. In other words, it does not utilize the fact that the motion and label fields in subsequent frames are very similar and that their relations can be formally expressed. Instead, it requires the computationally expensive re-estimation of a motion field. Furthermore, in the motion hypotheses extraction phase the spatial constraints are not fully utilized. In the block-matching motion estimation scheme (a) the block-based spatial constraints which are imposed fail near motion edges and (b) spatial constraints are not imposed between neighboring blocks. In the clustering phase spatial constraints are imposed only within intensity segments and not between neighboring intensity segments.

In this chapter we present an approach in which spatial and temporal constraints are incorporated into a single framework, allowing the joint estimation of the segmentation field and of the motion information[1]. The method operates at two levels (fig. 4.1). At the lower level (LEVEL 1 in fig. 4.1) a segmentation algorithm operating on the current frame's intensities provides a set of segments with relatively small intensity variation. Since the absence of an intensity edge is likely to imply the absence of a motion discontinuity, each segment is assumed to belong to a single object. Motion segmentation is then equivalent to grouping these segments into regions that move with the same motion parameters, that is assigning an "object" label to each segment. The number of the "object" labels is considered as known and are provided as a user-specified parameter. Finally, we assume that an affine parametric model is sufficient to

---

[1] The basis of this chapter appears in [93]

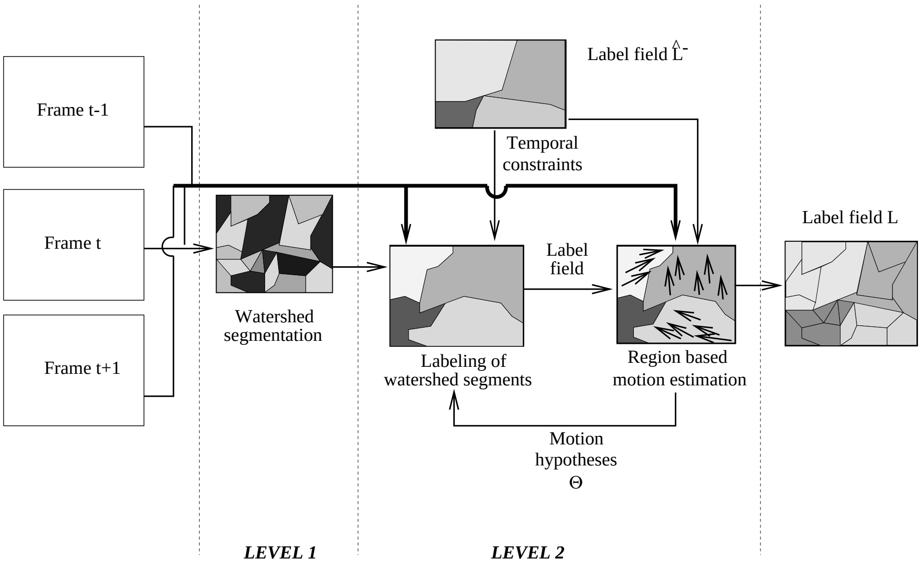


Figure 4.1: Outline of the joint motion estimation/segmentation approach

describe the various motion patterns in the scene.

The labeling is performed at the next level (LEVEL 2 in fig. 4.1) where the well-known notion of Markov Random Fields (MRF) is used in order to express spatial and temporal constraints on the level of the “intensity” segments. The criterion is the maximization of the conditional *a posteriori* probability (MAP) of the label field given the motion hypotheses, the label field of the previous frame and the image intensities. The equivalence between the MRF and the Gibbs distribution is exploited and the labeling is formulated as an optimization problem with respect to the motion parameters and the label field itself. For the optimization procedure we propose a method which reduces the corresponding objective function in an iterative way with respect to the motion parameters and the label field. In the optimization with respect to the motion parameters (*motion estimation phase*) the estimation is performed on region level, that is, a set of parameters is estimated for the collection of segments that comprise a single object. The estimation is constrained by the intensity preservation principle and by the temporal coherency of the label field. In the optimization with respect to the label field (*labeling phase*) an object label is assigned to each “intensity” segment. The labeling is constrained by the intensity preservation principle and the spatial and temporal coherency of the label field. A three-frame approach is taken in order to deal with occlusions.

In comparison to the method presented in chapter 3 the method proposed in this chapter also assumes parametric motion models, models the label field as a Markov Random Field and attempts a labeling based on the motion-compensated intensity differences. On the other hand, it addresses temporal issues and incorporates all of the constraints in a single framework for the **Joint** estimation of the motion hypotheses and the label field. That change of paradigm affects both the modeling and the corresponding optimization procedure.

The remainder of the chapter is organized as follows. In sections 4.1 and 4.2 we present the formulation of the labeling problem in the MAP framework and the optimization procedure, respectively. In section 4.3 experimental results are presented and finally conclusions are drawn in section 4.4.

4.1 Problem Modeling

We consider the supervised framework where the number of independently moving objects in the scene, denoted by N, is considered as known. We assume that the 2-D apparent motion field induced by them can be approximated by 6-parameter affine models. Affine motion models have been proved useful for motion estimation [37] [118], motion segmentation [18] [23] [114] and tracking [66] [98] and are a good compromise between efficiency and complexity. Keeping the terminology of chapter 3 we denote with $\theta = \{\theta(1) \ldots \theta(6)\}$ the parameters of the model and $\tilde{\mathbf{v}}_{\mathbf{i}}$ the model-generated motion vector at image location $\mathbf{i} = (\mathbf{i}_x, \mathbf{i}_y)$:

$$\tilde{\mathbf{v}}_{\mathbf{i}} = \left[\begin{array}{c} \theta(1)\mathbf{i}_x + \theta(2)\mathbf{i}_y + \theta(3) \\ \theta(4)\mathbf{i}_x + \theta(5)\mathbf{i}_y + \theta(6) \end{array} \right] \tag{4.1}$$

The motion hypotheses for the scene are the collection of the motion hypotheses for each object n, that is:

$$\Theta = \{\theta_n : n \in [1 \ldots N]\} \tag{4.2}$$

We seek for the unknown label field $L = \{l_s : l_s \in [1 \ldots N], s \in [1 \ldots K]\}$ and the motion hypotheses Θ at time instant t, considering as known the estimation of label field $\hat{L}^-$ at the previous time instant. We consider the Bayesian framework and more specifically we adopt as the labeling criterion the maximization of the *a posteriori* probability (MAP) of the label field. The conditional probability distributions to which the MAP criterion decomposes are modeled as Gibbs distributions. This MAP-MRF framework has been used extensively for regularization and for expressing contextual constraints in numerous problems in computer vision [42] [57].

More specifically, in the MAP framework, we aim for the maximization of the *a posteriori* probability:

$$P\left(L|I, \Theta, \hat{L}^-, I^-, I^+\right) \propto P\left(I|L, \Theta, \hat{L}^-, I^-, I^+\right) P\left(\hat{L}^-|\Theta, L\right) P\left(L|\Theta\right) \tag{4.3}$$

with respect to L and Θ. With I^-, I and I^+ we denote the image intensities in the previous, current and next frame respectively. We model appropriately the probability distribution functions in order to express the relations between the label fields, the intensities and the motion hypotheses. Due to the Hammersley-Clifford theorem, which states the equivalence between MRFs and Gibbs distributions (e.g. [10]), the modeling reduces to the definition of energy functions which are composed of local clique potentials.

The first term on the right-hand side of eq. 4.3 is the conditional probability distribution $P(I|L,\Theta,\hat{L}^-,I^-,I^+)$ which expresses how well the current motion and label field conform with the image intensities. We model it as a Gibbs distribution:

$$P(I|L,\Theta,\hat{L}^-,I^-,I^+) = \frac{1}{Z_1} e^{-E_d\left(I,L,\Theta,I^-,I^+\right)} \tag{4.4}$$

where Z_1 is the sum of the exponential term of eq. 4.4 over all possible realizations of intensities (all possible current frames). In general, Z_1 is dependent on the unknown fields L and Θ to the degree that the intensity variability along each trajectory is dependent on the intensity variability along other trajectories [56]. The independency assumption, although it is violated for example in occlusion areas, is one of the most common assumptions in the motion-related literature and will be also adopted in our work. Under this assumption, Z_1 can be considered as a normalization constant which does not effect the optimization procedure.

The energy term $E_d\left(I,L,\Theta,I^-,I^+\right)$ is formed by the Gibbs potentials V_{ds} as follows:

$$\begin{aligned} E_d\left(I,L,\Theta,I^-,I^+\right) &= \sum_{s=1}^{K} V_{ds}\left(I,s,\theta_{l_s},I^-,I^+\right) \\ &= \sum_{s=1}^{K} \min\left(\sum_{\mathbf{i}\in G_s}\left(f_{\mathbf{i}}^-\left(\theta_{l_s}\right)\right)^2, \sum_{\mathbf{i}\in G_s}\left(f_{\mathbf{i}}^+\left(\theta_{l_s}\right)\right)^2\right) \end{aligned} \tag{4.5}$$

where $f_{\mathbf{i}}^+\left(\theta_{l_s}\right)$ and $f_{\mathbf{i}}^-\left(\theta_{l_s}\right)$ are respectively the forward and backward motion-compensated intensity differences at pixel $\mathbf{i}$ ($\mathbf{i} \in G_s$). That is,

$$f_{\mathbf{i}}^+\left(\theta_{l_s}\right) = I(\mathbf{i}) - I^+(\mathbf{i} + \mathbf{v}_{\mathbf{i}}\left(\theta_{l_s}\right)) \tag{4.6}$$

$$f_{\mathbf{i}}^-\left(\theta_{l_s}\right) = I(\mathbf{i}) - I^-(\mathbf{i} - \mathbf{v}_{\mathbf{i}}\left(\theta_{l_s}\right)) \tag{4.7}$$

We make use of the intensity preservation principle, that is, we assume that the image intensities along the motion trajectories remain constant. The motion-compensated differences have the characteristics of the expected noise. Assuming that the noise is independent, the assumption that Z_1 in eq. 4.4 is constant is true.

Note that we are using a three-frame approach, where the motion-compensated intensity differences are defined either in the previous or in the next frame using the $\min$ operator. The direction (backward or forward) in which they are defined is common for the pixels that belong to the same intensity segment. By doing so, we are dealing in a simple and efficient way with appearing and disappearing areas. The underlying assumption is that these areas are visible in at least two consecutive frames, that is, pixels within each watershed segment have correspondences either in the next or in the previous frame. A similar approach where a visibility set for each pixel is defined in a forward/backward manner, is presented by Dubois and Konrad[32]. In their work the direction in which the motion estimation is constrained is determined by an occlusion field. In our segment-based approach the direct association of an occlusion field with the direction of the motion is not trivial since segments may be only partially occluded.

However, our assumption that the pixels within each segment have correspondences either in the next or in the previous frame, is violated only when a segment is large enough to partially appear **and** partially disappear from the scene. In this case the relative size and texture structure of the visible area will determine the accuracy of the temporal evidence for the segment in question. In such cases the spatial constraints need to provide the additional cues for correct labeling.

The second term on the right-hand side of eq. 4.3 expresses the temporal constraints. The conditional probability of the estimate of the label field at the previous frame $\hat{L}^-$ is modeled as a Gibbs distribution.

$$P\left(\hat{L}^-|\Theta, L, I^-, I^+\right) = \frac{1}{Z_2} e^{-E_t(L,\Theta,\hat{L}^-)} \tag{4.8}$$

where we consider Z_2 to be a constant following the same reasoning as for Z_1. The energy term $E_t(L, \Theta, \hat{L}^-)$ is formed by the Gibbs potentials V_{ts}, which are defined over single-site (single-segment) cliques as follows:

$$E_t(L, \Theta, \hat{L}^-) = \sum_{s=1}^{K} V_{ts}\left(\hat{L}^-, s, \theta_{l_s}\right) \tag{4.9}$$

The local energy term $V_{ts}\left(\hat{L}^-, s, \theta_{l_s}\right)$ expresses the cost (due to temporal constraints imposed on the label field) to assign label l_s to the watershed segment s, taking into consideration the motion hypothesis (θ_{l_s}) that this assignment implies. It is defined as

$$V_{ts}\left(\hat{L}^-, s, \theta_{l_s}\right) = z_t Q_s \tag{4.10}$$

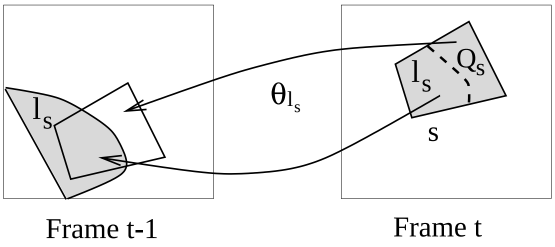


Figure 4.2: Temporal projection of watershed segment s (gray polygon in Frame t). The shaded area in frame $t-1$ represents a *region* labeled with label l_s

Each pixel of the watershed segment s is projected in the label field of the previous frame using the motion parameters θ_{l_s} (fig. 4.2). The number Q_s is the number of pixels of segment s whose motion-based projections in the previous frame have different labels. The smaller the Q_s, the higher the probability that s has the label l_s. The term z_t is a constant that controls the temporal consistency of the label field. The higher the z_t the higher the tendency to obtain a label field L and a set of motion hypotheses Θ such that each region n in the current frame is projected entirely within a region with the same label in the previous frame.

Finally, the third term on the right-hand side of eq. 4.3 models the *a priori* probability of the label field. We model it as a Gibbs distribution whose energy term $E_c(L)$ is formed by the Gibbs potentials V_c as follows:

$$E_c(L) = \sum_{s=1}^{K} V_{cs}\left(L, s\right) = \sum_{s=1}^{K} \sum_{s' \in N_s} V_c\left(s, s'\right) \tag{4.11}$$

The spatial Gibbs potentials $V_c\left(s, s'\right)$ are defined over pair-site (pair-segment) cliques as follows:

$$V_c\left(s, s'\right) = \begin{cases} -z_c b\left(s, s'\right) & \text{if } l_s = l_{s'} \\ z_c b\left(s, s'\right) & \text{if } l_s \neq l_{s'} \end{cases} \tag{4.12}$$

where z_c is a constant that controls the weight of the spatial constraints relative to the temporal constraints and to the constraints imposed by the intensity conservation principle. The term $b\left(s, s'\right)$ denotes the length of the common border between s and s'. We estimate this as the number of pairs of pixels $(\mathbf{i}, \mathbf{i}')$ which are neighbors in the image grid and belong to the borders of s and s' respectively. The maximization of $P\left(L|\Theta\right)$ is equivalent to the minimization of $E_c(L)$. We add a negative quantity to the potential $E_c(L)$ whenever two neighboring segments have the same label, and a positive quantity when they have a different label. In this way we favor configurations of the label field that are spatially smooth. This quantity, which can be interpreted as the spatial interaction between watershed segments, is proportional to $b\left(s, s'\right)$, in other words, their spatial interaction increases with the length of their common border. From a global point of view, a minimization with respect to the spatial energy term $E_c(L)$ tends to minimize the total border length between neighboring objects.

The parameters z_c and z_t control the relative weights of the spatial and the temporal constraints in the estimation of the label field, with respect to the data energy term $E_d(I, L, \Theta, I^-, I^+)$. For setting their values one needs to understand the relative importance of the three energy terms. As far as the local data energy $V_{ds}\left(I, s, \theta_{l_s}, I^-, I^+\right)$ and the local temporal energy $V_{ts}\left(\hat{L}^-, s, \theta_{l_s}\right)$ are concerned, both of them are estimated over the ensemble of the pixels of the segment s. That is, each pixel contributes to the local energy on the one hand with the square of the backward/forward motion-compensated intensity difference and on the other hand with z_t in the case that its label is different than the label of its motion-based projection in the previous frame. In order to achieve a balance between these two terms, z_t should be set according to the expected characteristics of the motion-compensated intensity differences. However, the correctness of the influence of the temporal constraints depends on the degree of the accuracy of the estimated label field $\hat{L}^-$ in the previous frame. To ensure that the algorithm is flexible enough to correct errors in $\hat{L}^-$, the value of z_t should be chosen rather conservatively. In all of our experiments the value of z_t was chosen in the range between 1 and 2.

Since the data energy term and the temporal energy term are estimated over the same ensemble, that is, the ensemble of pixels, a natural normalization in the contribution of these terms is achieved. On the other hand, the spatial energy term is defined in terms of the lengths of the common borders between neighboring segments. At segment level, the influence of the data energy term and the temporal energy term is proportional to the number of pixels of segment s, while the influence of the spatial energy term is proportional to the perimeter of s. In general, this implies that the larger a segment, the larger the ratio between the relative contribution of the local data and temporal energy terms with respect to the local spatial energy term. This marks a gradual shift on the weight we assign to the different sources of evidence that we utilize for the labeling. For larger segments emphasis is placed on the evidence that the segments themselves provide about their temporal behavior while for smaller segments the emphasis is placed on the evidence provided by the label field in their neighborhood.

For the manual setting of the value of z_c a rule of thumb can be provided by an analysis of the ratio between the size and the perimeter of the segments. Such an analysis can provide the value of z_c for which a normalization between the three energy terms is approximately achieved for segments of a given size. This can be useful when the size of the smaller expected object is *a priori* known. The relation between the size and the perimeter of the segments depend on two factors. On the one hand it depends on the shape of the segments with the two extremes being a) very elongated segments that yield a linear relation and b) circular segments that yield a quadratic relation with the largest possible coefficient ($(4\pi)^{-1}$). On the other hand, it depends on the definition of the perimeter. For our definition of the border length and for relatively small segments (less than 150 pixels) an approximation of the segment's size a as a function of its perimeter p was experimentally found to be $a = \frac{1}{36}p^2$. This implies that a normalization of the different energy terms for a segment with 100 pixels is achieved for a value of 1.5 for z_c. However, if we take into consideration that the data energy term has the characteristics of the variance of the motion-compensated intensity differences and that the temporal energy term is scaled by the factor z_t, a value of z_c in the range between 3 and 10 is a more reasonable choice.

The parameter z_c can be also interpreted as a smoothing parameter, and in this context its automatic determination has received a lot of attention, especially as a regularization parameter for the solution of ill-posed problems [9] [38]. In our case, a recently presented voting technique [90] could be applicable for simultaneously estimating the label field and the parameters z_c. That approach, based on an estimation of the current label field, attempts to obtain a better estimation of the interaction parameter in terms of the local characteristics of the energy function. Since the number of sites (watershed segments) is much lower than that of pixel-based approaches the computational overhead would be greatly reduced in comparison to them. However, issues such as the degree of dependence on the quality of the initial label field and the applicability of the method for the automatic determination of z_t should be further investigated.

As far as the assumption that the number of objects N is concerned, we should note that the maximization of the *a posteriori* probability that we used as a criterion might be also suitable for the automatic determination of N. Under our modeling assumptions for the probability distributions of the intensity and the label field of the current frame, the sum of the three energy terms is proportional to the encoding cost of the current frame if we use an optimal code. This stems from the fact that the sum of the energy terms is proportional to the sum of the negative logarithms of eq. 4.4 and eq. 4.8, that is, the sum of the encoding cost for the intensity residual and the label field respectively. In the above analysis we have ignored the cost of encoding the motion parameters as this cost can be considered negligible in comparison to the other terms. Practically, a minimization with respect to the number of objects N is a compromise between the accuracy of the prediction of the intensities at the current frame and the complexity of the label field. More specifically, when we add a new label we reduce the data energy term E_d. This comes at the cost of a most probable increase of the spatial energy term E_c, which depends on the common border length between the different objects, and a definite increase of the temporal energy term E_t, which penalizes label fields which are not similar to the one estimated in the previous frame.

The automatic determination of the number of objects using the Minimum Description Length principle [96] [111] has often been addressed in the literature (e.g. [97] [27]). Following the most commonly used scheme we could apply our method for the certain values of N and choose the one which at convergence yields the lowest sum of energies. However, since we have not conducted any experiments in that direction, the issue of the automatic determination of N should be considered in the context of future work.

4.2 MAP estimation

Once the energy functions are defined, the MAP estimate is equivalent to the minimization of the quantity

$$E(L,\Theta,I,\hat{L}^-,I^-,I^+) = E_d(I,L,\Theta,\hat{L}^-,I^-,I^+) + E_c(L) + E_t(L,\Theta,\hat{L}^-) \tag{4.13}$$

The minimization of eq. 4.13 is a non-linear optimization problem with respect to L and Θ. For solving such optimization problems, a number of methods have been proposed in the literature which range from deterministic [11][25] and stochastic[53][42] relaxation schemes to methods inspired by Genetic Algorithms[58]. Stochastic optimization can provide better solutions in terms of the objective function due to its ability to escape local minima, but it is computationally intensive. Deterministic relaxation methods on the other hand, can provide a good solution to a local minimum if presented with a reasonable initialization. Since in our case initializations are available from the estimations made in the previous frame we will restrict ourselves to the deterministic framework. We propose a method which iterates between a minimization with respect to the label field (labeling phase) and a minimization with respect to the motion parameters (motion estimation phase). In the labeling phase a relaxation algorithm is employed to solve the combinatorial problem of assigning object labels to watershed segments. In the motion estimation phase eq. 4.13 is linearized with respect to Θ and a gradient-based approach is adopted. Let us note that both minimization phases employ iterative algorithms. In order to avoid confusion we will refer to the iterations between the motion estimation and the labeling phase as *external* and to the iterations within each phase as *internal*.

More specifically, let us denote with m the index for the external iterations. Then the MAP estimation at the external level iterates between the following phases:

$$L_{m+1} = \arg\min_{L} E(L, \Theta_m, I, \hat{L}^-, I^-, I^+) \tag{4.14}$$

$$\Theta_{m+1} = \arg\min_{\Theta} E(L_{m+1}, \Theta, I, \hat{L}^-, I^-, I^+) \tag{4.15}$$

The interaction between the external and the internal iterations is depicted in fig. 4.3, where with k we denote the index for each of the internal iterations. At the level of the internal iterations the index m is omitted for the sake of notational simplicity. When convergence is achieved at one of the internal levels (motion estimation or labeling), or after a fixed number of iterations, the control is transferred to the other internal level (labeling or motion estimation respectively). In our implementation we used a fixed number of iterations for the motion estimation phase.

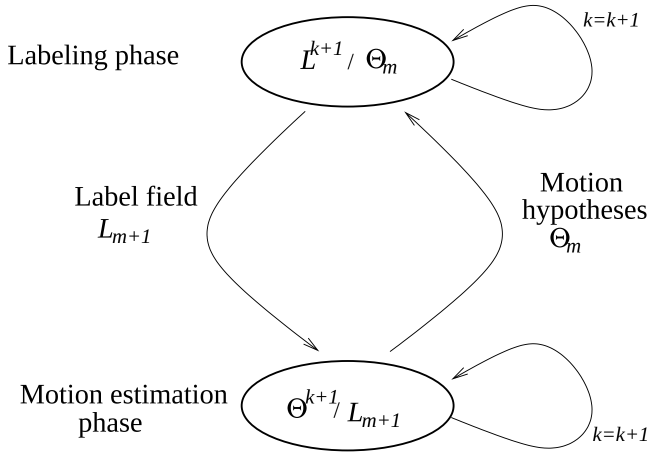


Figure 4.3: Flow diagram of the optimization procedure. Motion hypotheses extraction and labeling are performed in an iterative way.

For the first external iteration (i.e. for $m = 0$) the motion hypotheses are initialized with the motion parameters estimated for the previous frame.

Let us note that the optimization procedure described by eq. 4.14 and eq. 4.15 bears similarities to the *Expectation Maximization* method, when hard classification is employed. In that framework, in the *Expectation* step the Expectation of the conditional probability of each of the labels at each site given the estimate of Θ at iteration m should be estimated. The label with the highest conditional probability is then assigned to the site in question. However, the estimation of the Expectation of the conditional probabilities of the latent data (i.e. the label field) is complicated when the mode introduces interdependencies in the label field. Zhang and Modestino [120] generate samples of the label field through simulated annealing and compute the conditional expectations by Monte Carlo when the interdependencies are expressed using a Markov Random Field modeling. Weiss and Adelson [115] also use the label field as a Markov Random Field but based on results which relate the EM algorithm to statistical physics

[78] validate guesses of the conditional expectations in the *Expectation* step. In the case that the guesses fail to reduce the free energy they employ gradient updates to the old estimates of the conditional expectations. In our case, each internal iteration in the labeling phase which we employ could be interpreted as locally updating the conditional expectations based on the label field estimated in the previous internal iteration and choosing the label with the highest conditional probability. Since the choice of the label with the highest conditional probability implies a reduction of the total energy (Eq. 4.13) the convergence of the scheme is guaranteed. In appendix C we show that our method is essentially an EM method with hard decisions and that with a minor modification it can incorporate soft decisions too.

4.2.1 Labeling phase

In the labeling phase (eq. 4.14) the minimization of eq. 4.13 with respect to L takes place, keeping the motion parameters Θ "frozen" ($\Theta = \Theta_m$). We employ an iterative deterministic relaxation algorithm known as Iterative Conditional Modes (ICM). Proposed by Besag [11], ICM maximizes iteratively the conditional probability of a label at each site, given the labeling at all other sites. In the original algorithm at each iteration each site is visited and assigned the label that maximizes that conditional probability. Due to the local interactions that the MRF modeling implies this reduces to the calculation of the local clique potentials for each site under consideration.

As stated in [11] the order in which sites are visited is important for the final configuration. Furthermore, a site can contribute to the reduction of the energy only if the local labeling configuration has changed, that is, it is not necessary to visit all sites at each iteration. In order to cope with the latter we maintain a set of candidate segments c_k for each iteration k and in order to avoid the influence of a predetermined ordering we adopt a random visit schedule on the elements of the candidate set. The steps of the algorithm are summarized in Table 4.1.

For the estimation of the label field L in the first internal iteration of the first external iteration we take into account only the motion hypotheses, and the temporal constraints, that is, we minimize the quantity $E_d(I, L, \Theta, I^-, I^+) + E_t(L, \Theta, \hat{L}^-)$. No spatial interactions are introduced at that point, that is, the label of a watershed segment s in independent of the labels of the rest of the watershed segments. This serves as an initialization of the label field for the optimization procedure in the current frame.

4.2.2 Motion Estimation phase

In the motion estimation phase the minimization of eq. 4.13 with respect to Θ takes place, keeping the label field L "frozen" ($L = L_m$). This minimization is a non-linear optimization problem since neither $E_d\left(I, L, \Theta, I^-, I^+\right)$ nor $E_t(L, \Theta, \hat{L}^-)$ are linear with respect to the motion parameters. For $E_d\left(I, L, \Theta, I^-, I^+\right)$ this is because firstly the image intensities are non-linear with respect to Θ and secondly because of the non-linear minimum operator (eq. 4.5). In order to overcome the latter we turn the minimization of eq. 4.13 into an equivalent optimization problem by introducing a binary **direction** field $\{d_s : d_s \in [0, 1], s \in [1 \ldots K]\}$. This field determines the

```
1. $c_0 = \{s : s \in [1 \dots K]\}, c_1 = \emptyset, k = 0$
2. Choose randomly a segment $s$ from $c_k$
3. Assign to $s$ the label $l$ that minimizes the local energy:
       $l_s = \arg\min_l \left( V_{ds}\left(I, s, \theta_l, I^-, I^+\right) + V_{cs}\left(L, s\right) + V_{ts}\left(\hat{L}^-, s, \theta_l\right) \right)$
4. $c_k = c_k - \{s\}$
5. If $s$ has changed label, update the candidate list for the next iteration
       $c_{k+1} = c_{k+1} \cup \{s' : s' \in N_s\}$
6. If $c_k \neq \emptyset$ go to step 2
7. If $c_{k+1} \neq \emptyset$ then $k = k + 1$ and go to step 2. Otherwise STOP
```

Table 4.1: Modified ICM for segment labeling

direction, backward or forward, in which the temporal intensity variation constrains the motion estimation. Let us define $C_e\left(d, \Theta\right)$ as:

$$\begin{aligned} C_e\left(d, \Theta\right) = {} & E_t(L, \Theta, \hat{L}^-) + \\ & \sum_{s=1}^{K} \left(d_s \sum_{\mathbf{i} \in G_s} \left(f_{\mathbf{i}}^+\left(\theta_{l_s}\right)\right)^2 + (1 - d_s) \sum_{\mathbf{i} \in G_s} \left(f_{\mathbf{i}}^-\left(\theta_{l_s}\right)\right)^2 \right) \end{aligned} \tag{4.16}$$

where the functional dependence of $C_e\left(d, \Theta\right)$ on the label fields and on the image intensities is omitted for notational simplicity. The new energy term $C_e\left(d, \Theta\right)$ is derived from the terms of eq. 4.13 that depend on Θ.

Lemma 1
If $(\hat{\Theta}, \hat{d})$ are the arguments which minimize eq. 4.16 then $\hat{\Theta}$ is the argument which minimizes eq. 4.13 with respect to Θ.

Proof is given in appendix A

We minimize eq. 4.16 with a method which iterates between a minimization with respect to Θ and a minimization with respect to d. More specifically:

$$d^{k+1} = \arg\min_d C_e\left(d, \Theta^k\right) \tag{4.17}$$

$$\Theta^{k+1} = \arg\min_\Theta C_e\left(d^{k+1}, \Theta\right) \tag{4.18}$$

where k is the iteration index. Θ^0 are the motion hypotheses estimated in the previous external iteration or, for the first external iteration, the motion hypotheses estimated for the previous frame.

Clearly the minimization of $C_e\left(d, \Theta^k\right)$ with respect to d yields:

$$d_s = \begin{cases} 1 & \text{if } \sum_{\mathbf{i} \in G_s} \left(f_{\mathbf{i}}^{+}\left(\theta_{l_s}\right)\right)^2 \leq \sum_{\mathbf{i} \in G_s} \left(f_{\mathbf{i}}^{-}\left(\theta_{l_s}\right)\right)^2 \quad \text{(forward)} \\ 0 & \text{otherwise} \quad \text{(backward)} \end{cases} \tag{4.19}$$

For the minimization of $C_e\left(d^{k+1}, \Theta\right)$ with respect to Θ we have available the field d^{k+1}, which gives us for each segment s the direction (backward or forward) in which motion information has to be considered. Let us stress the fact that **one** parametric model is estimated per object, but that for each of the segments constituting the object the direction might differ. In order to solve for Θ we first make a linear approximation of $C_e\left(d^k, \Theta\right)$ with respect to Θ. We will solve for Θ in an incremental way which iteratively improves the previous estimation Θ^k. Let us express the motion parameters Θ as $\Theta = \Theta^k + \Delta\Theta$. At iteration $k+1$ we will solve for the $\Delta\Theta$ [83] for which the gradient of C_e with respect to the motion parameters is zero.

First we make a linear approximation of the image intensities with respect to the motion parameters by expanding $f_{\mathbf{i}}^{-}(\theta)$ (or $f_{\mathbf{i}}^{+}(\theta)$) using a first-order Taylor approximation [47][83]. More specifically suppose that for a segment s it is the case that $d_s = 0$. Then at each point $\mathbf{i} \in G_s$:

$$\begin{aligned} f_{\mathbf{i}}^{-}\left(\theta_{l_s}\right) &= I(\mathbf{i}) - I^{-}(\mathbf{i} - \tilde{\mathbf{v}}_{\mathbf{i}}(\theta_{l_s})) \\ &\approx I(\mathbf{i}) - \left(I^{-}(\mathbf{i} - \tilde{\mathbf{v}}_{\mathbf{i}}(\theta_{l_s}^k)) - \nabla I^{-}(\mathbf{i} - \tilde{\mathbf{v}}_{\mathbf{i}}(\theta_{l_s}^k))\tilde{\mathbf{v}}_{\mathbf{i}}(\Delta\theta_{l_s})\right) \end{aligned} \tag{4.20}$$

eq. 4.20 is linear with respect to $\Delta\theta_{l_s}$ thus the second term on the right-hand side of eq. 4.16 (i.e. the summation) is now approximated by a summation of the squares of terms like eq. 4.20. Therefore, its gradient is linear with respect to $\Delta\theta_{l_s}$. We deal with subpixel accuracies using a bicubical interpolator which is continuous in its first derivatives [52].

In order to be able to differentiate eq. 4.16 with respect to the motion parameters we will also need an approximation of the term $E_t(L, \Theta, \hat{L}^{-})$. For each object n (where $n \in [1 \ldots N]$) we define a field O_n^{-} such that:

$$O_n^{-}(\mathbf{i}) = \begin{cases} 0 & \text{if } \hat{L}_{\mathbf{i}}^{-} = n \\ \sqrt{z_t} & \text{otherwise} \end{cases} \tag{4.21}$$

where $\hat{L}_{\mathbf{i}}^{-}$ denotes the label of the segment to which the point $\mathbf{i}$ belongs. Note that the field O_n^{-} is defined in terms of the label field estimated for the previous frame.

Let us express $E_t(L, \Theta, \hat{L}^{-})$ in terms of O_n^{-}:

$$\begin{aligned} E_t(L, \Theta, \hat{L}^{-}) &= \sum_{s=1}^{K} V_{ts}\left(\hat{L}^{-}, s, \theta_{l_s}\right) \\ &= \sum_{s=1}^{K} \sum_{\mathbf{i} \in G_s} \left(O_{l_s}^{-}(\mathbf{i} - \tilde{\mathbf{v}}_{\mathbf{i}}(\theta_{l_s}))\right)^2 \end{aligned} \tag{4.22}$$

In order to approximate eq. 4.22 with a function which is quadratic with respect to the motion parameters, we smooth O_n^- with a Gaussian filter with a small variance and, as was the case for the intensities, we consider the first-order Taylor approximation. We will express eq. 4.22 in a way that allows an incremental motion estimation. More specifically:

$$E_t(L, \Theta, \hat{L}^-) \approx \sum_{s=1}^{K} \sum_{\mathbf{i} \in G_s} \left(O_{l_s}^-(\mathbf{i} - \tilde{\mathbf{v}}_{\mathbf{i}}(\theta_{l_s}^k)) + \nabla O_{l_s}^-(\mathbf{i} - \tilde{\mathbf{v}}_{\mathbf{i}}(\theta_{l_s}^k)) \tilde{\mathbf{v}}_{\mathbf{i}}(\Delta\theta_{l_s}) \right)^2 \tag{4.23}$$

Under the approximations of eq. 4.20 and eq. 4.23 the energy function $C_e\ (d, \Theta)$ is quadratic with respect to the motion parameters. Expressing Θ as $\Theta = \Theta^k + \Delta\Theta$, the correction in the motion parameters $\Delta\Theta$ is given by solving:

$$\nabla C_e \left(d^k, \Theta^k + \Delta\Theta \right) = 0 \tag{4.24}$$

where ∇C_e is the gradient of C_e with respect to the motion parameters. eq. 4.24 results in N linear systems with six unknowns: one for each of the N sets of motion parameters θ_n.

4.3 Experimental Results

We have applied the proposed method in a number of image sequences in order to test the validity of our approach. We present results for three sequences in each of which different challenges arise. The first one is the MPEG-4 validation sequence “coast guard”. In this image sequence we are presented with a complex scene, in which four different objects are present. The motions of the objects are quite small in magnitude which makes the distinction between them rather difficult. In the second one, the “train” sequence, we are again presented with a complex scene with three different apparent motion patterns. In this case the motions are large, a fact which generates strong occlusions and even blurs the edges of one of the objects. In order to demonstrate the contribution of the direction field *d*, we present results obtained by considering only the forward direction, both in the motion estimation and in the labeling phase. Finally in the third sequence, the “pig” sequence, we are presented with an independently moving object in a static background. The scene is simple but the moving object exhibits small deformations in shape, and large rotational components are present in the motion pattern.

4.3.1 “coast guard” sequence

For the MPEG validation sequence “coast guard”, fig. 4.4(a) and fig. 4.4(b) depict an original frame and the corresponding validation label field which is used only for illustrative purposes. Given this labeling mask four different objects are present, namely the “water”, “left ship”, “middle ship” and the “shore”. The camera follows the ship in the middle, while another ship is entering the scene. The water of the river globally

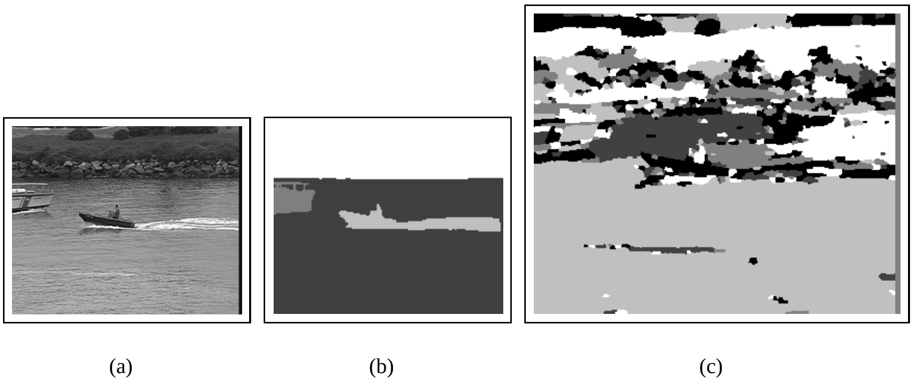

(a) (b) (c)

Figure 4.4: The 14th frame of “coast guard” sequence, the corresponding validation label field and the watershed segmentation

appears to move to the right, but deviations from the dominant motion pattern occur locally. The motion behavior of the different objects is quite similar; a distinction between the “shore” and the “water” is possible only at subpixel level.

The result of the watershed segmentation is depicted in fig. 4.4(c), where an area with constant intensity represents a watershed segment. Our primary goal of obtaining a well-localized, edge preserving segmentation is largely achieved. Except of the thin mast of the ship entering the scene, each watershed segment belongs entirely to a single object, a result which validates our choice of using a small structuring element.

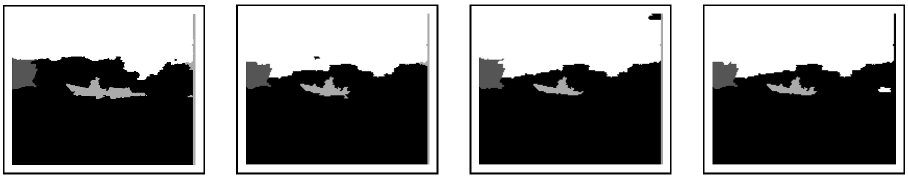


Figure 4.5: Label fields for the 10th frame of the “coast guard” sequence at external iterations 1, 2, 3 and 5

The initialization of the motion parameters for the 10th frame were obtained by an estimation based on the robust clustering described in chapter 3. Almost identical results were obtained by setting setting manually the horizontal translational components of the motion models (i.e. parameter a_3) to 1, 1.8, -0.5 and 1.3 for the “water”, “left boat”, “middle boat” and the “shore” respectively. In fig. 4.5 we present the label fields obtained at the end of successive external iteration, with the temporal constraints disabled. The algorithm is capable of distinguishing the different objects in the scene by grouping successfully the watershed segments into regions that move in the same way. Both of the ships are well localized and the “water” is separated from the “shore” except for the part which is closer to the shore. The latter is probably due to the difference in depth between the parts of the water that are closer to the camera and the parts that are closer to the shore. The main difference with the “ground truth” segmentation of fig. 4.4(b) remains the trail of the ship in the water, which in our case was merged with the big “water” segment. However, it is questionable, without any semantic reasoning, whether it is possible to classify with the same label the ship and a trail whose apparent movement is quite arbitrary. In comparison, we have disabled the spatial constraints and applied our method with the same initializations. The resulting label field is depicted in fig. 4.6. In the absence of spatial constraints, the motion of the “water” cannot be estimated well, a fact which results in misclassifications and in a “noisy” label field.

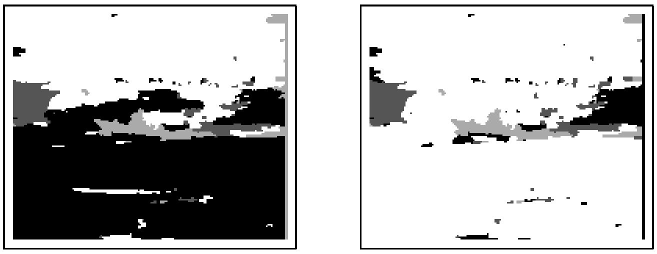


Figure 4.6: Label field for the 10th frame of “coast guard” sequence, without spatial and temporal constraints. Label fields are presented at external iterations 2 and 4

In order to obtain a good separation of the “shore” and the “water” we enabled the temporal constraints using a label map for frame 9, which was obtained by setting manually the motion parameters and applying only the labeling phase. Then, for frame 10, no initialization was used for the motion parameters (i.e. they were all set to zero). The label field which is used for frame 9 and the final label field obtained for frame 10 are presented in fig. 4.8 and the evolution of the horizontal motion component during the internal iterations is presented in fig. 4.7.

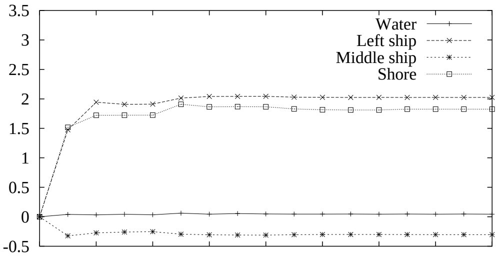


Figure 4.7: Evolution of the horizontal translational components ($\theta(3)$) of the affine motion model for the 10th frame of the “coast guard” sequence

In order to illustrate the temporal behavior of the method, we present in fig. 4.9 the original frames and the corresponding magnifications of the masks obtained for two

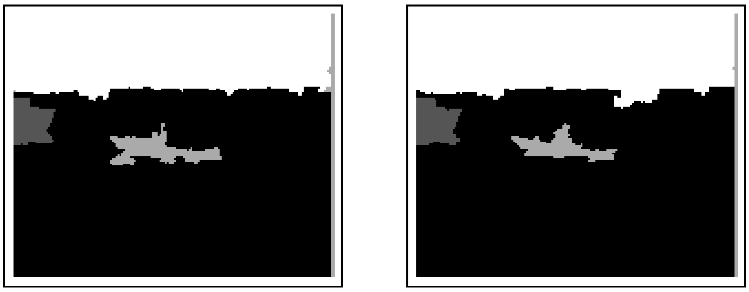


Figure 4.8: Label field used for the 9th frame of “coast guard” sequence and label field obtained for the 10th frame using spatial and temporal constraints

ships for frames 10 to 30, with a step of 10 frames. The method exhibits good temporal stability, and the edges are well localized in the subsequent label fields. In fig. 4.10 we present with a step of 5 frames, the label field and the corresponding mask of the “shore” object.

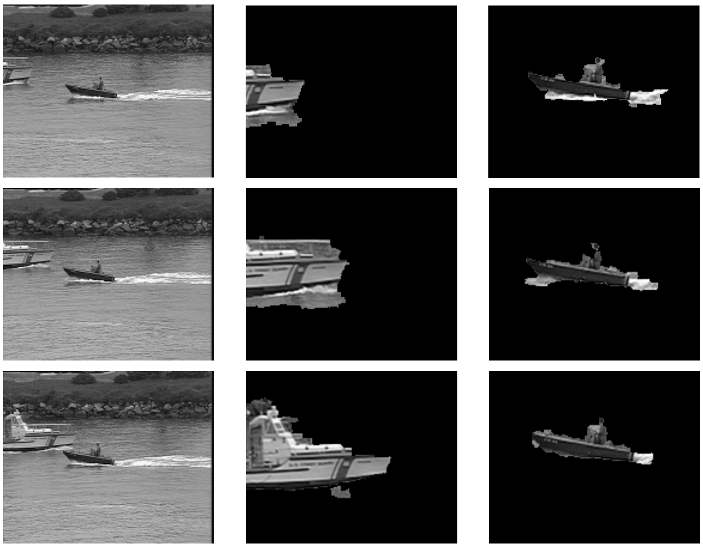


Figure 4.9: Original frames 10, 20 and 30 of the “coast guard” sequence with the corresponding magnified masks of the two ships

Finally, we have applied our method assuming that only three objects are present in the scene. We manually initialized the horizontal component of the motion parameters of the objects “left ship”, “background” and “middle ship” to 3, 0.1 and 0 respectively. In fig. 4.11(a) we present the label field obtained at convergence for the 10th frame of the sequence when both temporal and spatial constraints are disabled. In the absence of temporal and spatial constraints we could not obtain a good localization of the two ships, even though the results of fig. 4.6 indicated that this might be the case. For comparison, we present in fig. 4.11(b) the label field obtained for the same frame and with the same initializations but with the spatial constraints enabled. Finally, in fig. 4.11(c) we present the label field obtained at convergence for frame 11, with both the temporal and spatial constraints enabled. For this frame we used as $\hat{L}^{t-1}$ the label field of fig. 4.11(b).

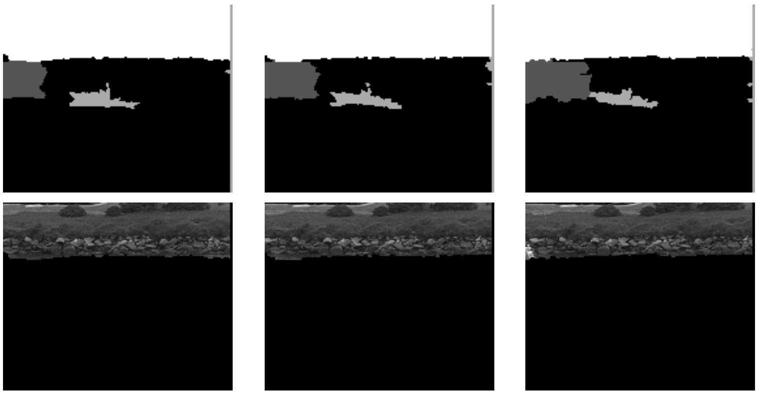
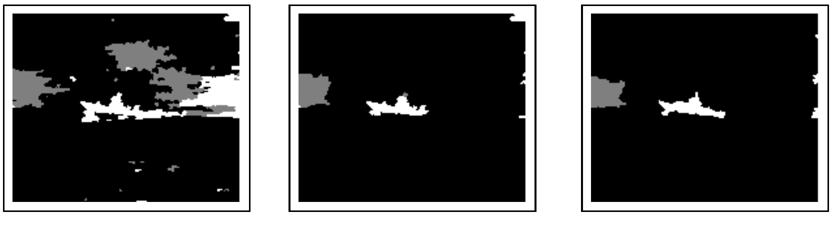


Figure 4.10: Label masks and “shore” segment for frames 15, 20 and 25 of the “coast guard” sequence



(a) (b) (c)

Figure 4.11: Label fields under the assumption that three objects are present in the scene. (a) Label field for the 10th frame without spatial and temporal constraints. (b) Label field for the 10th frame with spatial constraints. (c) Label field for the 10th frame with both spatial and temporal constraints

4.3.2 “train” sequence

The algorithm has been also tested on the even field of the interlaced “train” sequence. The original fields for frames 10 to 33 with a step of 8 frames are presented in the left column of fig. 4.12. The movement of the camera generates an apparent background motion of about 4 to 8 pixels per frame (depending on the relative depth), one train is moving with 6 pixels per frame and the other train with about 45 pixels per frame. Due to the large apparent motion large areas appear and disappear from the scene: the areas in front and behind the second train as well as areas at the borders of the image. For the same reason, there are areas that even appear only for one frame, for example the area between the wagons of train in the foreground (“train two”).

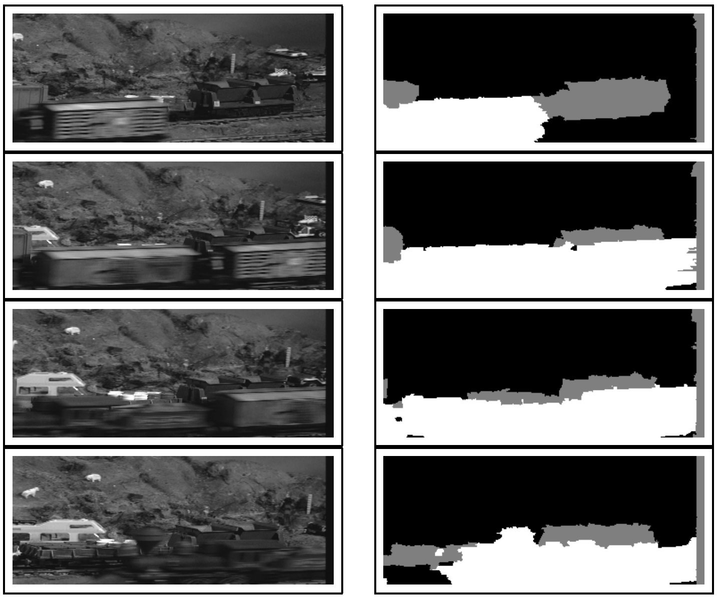


Figure 4.12: Original frames 10-37 of the “train” sequence with the corresponding label masks

We present results for 24 frames of this sequence, namely the frames 10 to 33, with a step of 8 frames. In fig. 4.12 the original frames and the corresponding label fields are depicted. In fig. 4.13 the masks for the two trains and the background are also presented. The algorithm exhibits good localization properties and the areas that appear and disappear are also classified successfully due to the bidirectional way in which we validate the motion hypotheses. Problems occur mainly in the areas between the wagons of the train in the foreground that appear only for one frame. Since there is no correspondence in the previous or in the next frame and the temporal constraint is also invalid, these areas are likely to be misclassified. The values of z_c and z_t were manually set to 4.5 and 2 respectively and the algorithm was applied considering frame 10 as the first frame of the sequence. Results for frame 10 were obtained with a rough manual initialization of the horizontal motion components of the 3 different objects to 35, 5 and -5 pixels/frame. For that frame, the temporal constraints were disabled since the label field of the previous frame was not available and the value of z_c was set to 6.5. The corresponding label field for the same initialization of the motion parameters and with the spatial and temporal constraints disabled is depicted in fig. 4.14.

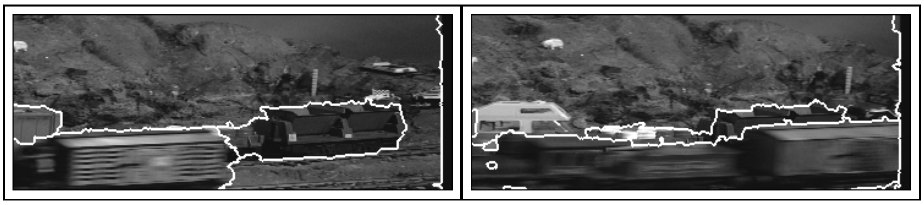


Figure 4.13: Object edges superimposed on the original frames 10 and 26 of the "train" sequence

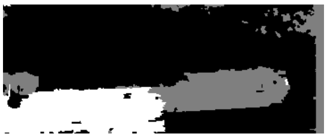


Figure 4.14: Label field for train sequence obtained without spatial and temporal constraints

In order to demonstrate the influence of the bidirectional way in which the motion hypotheses are estimated and validated, we present experimental results which were obtained by setting $\{d_s = 1, s \in [1 \ldots K]\}$. In this way only the forward direction is considered. In fig. 4.15 we present the edges of the corresponding label field superimposed on the 10th frame of the sequence. This is to be compared with the results presented in fig. 4.12 and fig. 4.13. Misclassifications occur in the areas that are covered in the next frame (frame 11), namely the areas in front of the two trains as well as in the right edge of the field of view.

In order to illustrate the internals of the iterative procedures and to provide insight in what degree the method depends on the initializations, we have applied the algorithm at the 10th frame of the sequence with a bad initialization for the motion parameters and have disabled the temporal constraints. The horizontal motion components of the 3 different objects were manually initialized to 13, 0, and 0.2 ("train 2", "background" and "train 1" respectively) and the parameter z_c was manually set to 4.3. A worse initialization of the motion parameters or a higher value of z_c causes the "train 2" and the "background" to merge. In general, our experiments indicated that when the motion parameters are badly initialized it is better to choose a lower value for the parameter z_c. However, even with these marginal initializations we were able to obtain a reasonably good estimate of the label field and of the motion parameters. In fig. 4.17 we present the label fields obtained at different stages of the external iterations and in fig. 4.16 we

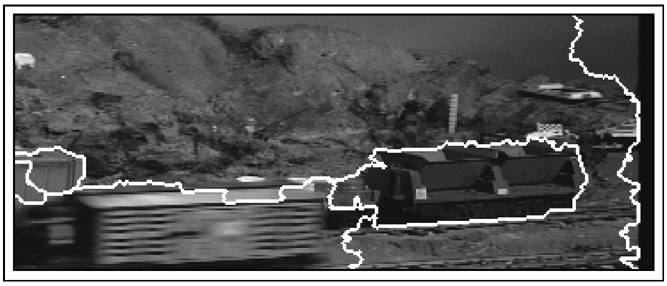


Figure 4.15: Object edges superimposed on the 10th frame of the “train” sequence with forward motion estimation and labeling $(\{d_s = 1, s \in [1 \ldots K]\})$. The absence of a correct match in the next frame causes misclassifications in occluded areas.

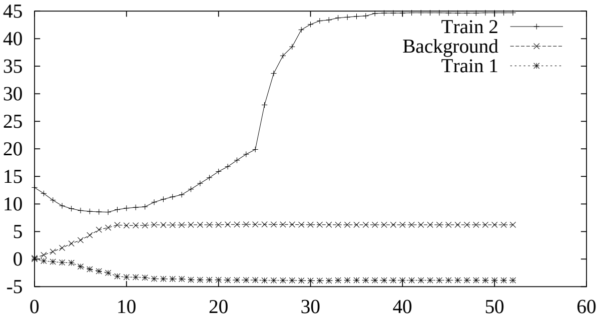


Figure 4.16: Evolution of the horizontal translational components ($\theta(3)$) of the affine motion model for the 10th frame of the “train” sequence

present the horizontal motion components obtained at subsequent applications of the motion estimation at the internal iterations. We should note that the motion parameters that we have obtained in this way are a much better initialization than the one we used to obtain the results of fig. 4.12.

4.3.3 “pig” sequence

In the “pig” sequence, which was obtained for the needs of a project for monitoring animal behavior, a pig is moving against a static background, with slowly changing illumination conditions. There is strong rotation in some of the frames of the sequence and deformations of the body of the pig. Moreover, the assumption of rigid motion is violated in areas like the pig’s ears and legs. In fig. 4.18 the label fields for frames 411, 416, 421 and 426 are presented. The localization accuracy and the temporal stability are preserved, even though the motion of the pig changes quite fast and in a strong rotational sense. However the motion of the ears and the leg, in some cases, deviate significantly from the estimated parametric model and merge with the background.

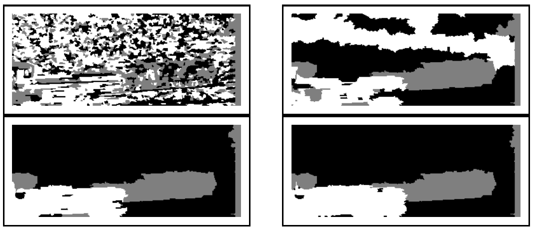


Figure 4.17: Label fields at subsequent external iterations with a bad initialization of the motion parameters and disabled temporal constraints. Top left: Maximum Likelihood labeling (without spatial and temporal constraints). Top right: Label field at the 3rd external iteration. Bottom left: Label field at the 6th external iteration. Bottom right: Final label field (12th iteration)

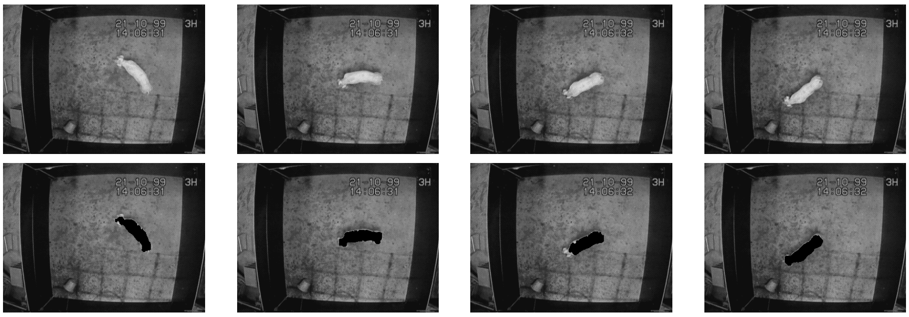


Figure 4.18: Original frames and masks for frames 411, 416, 421, 426 of the “pig” sequence (Courtesy of R. Griekspoor)

4.4 Conclusions

In this chapter we proposed a method for motion-based segmentation of video sequences, in which spatial and temporal consistency is expressed in terms of interactions between segments that result from an initial intensity segmentation. As a means for the initial intensity segmentation we used a watershed segmentation algorithm. Using a small structuring element for the morphological simplification of each frame we were able to preserve the significant edges and to achieve high localization accuracy.

We have defined in a formal way the spatial and temporal interactions between the watershed segments by building a model for the label field based on the notion of the Markov Random Fields. We express the labeling problem as an optimization problem with respect to the *a posteriori* probability of the label field. We propose a three-frame approach in order to deal with occlusions and develop an iterative optimization scheme. The relation of our method with pixel-based approaches has been demonstrated by considering the degenerate case, where the intensity segments contain a single pixel.

We have presented results for various image sequences, which show that with the proposed modeling it is possible to group segments that result from a “fine” initial segmentation, based on motion information. The proposed method exhibits good localization properties, temporal stability and deals successfully with motion occlusions.

For future work, an explicit treatment of the occlusions and more specifically of occlusions in the previous frame could be beneficial. This implies the identification of segments that have just appeared in the scene and the relaxation of the assumption of the temporal continuity of the label field in such cases. The direction field could provide the basis for analysis in that direction.

The relaxation of the rigidity assumption that is imposed by the affine parametric model would be an interesting direction to follow. In the motion-based segmentation framework such a relaxation usually involves the assumption that the motion field varies smoothly within each object and that object edges coincide with motion discontinuities. In our paradigm (i.e. a model applied on a “fine” initial intensity segmentation) such an extension is not a big step.

Finally, the automatic determination of the number of the objects in the directions that we have outlined in section 4.1 might be an interesting extension.

Acknowledgment

The video material for the “pig” sequence was provided by the Dept. of Animal Health and Welfare, Danish Institute of Agricultural Sciences. The results for the same sequence were obtained at Noldus Information Technology b.v., Wageningen, The Netherlands.

Chapter 5

Semi-automatic Statistical Segmentation

In the last two chapters we presented methods which discriminate between the objects that compose the scene on the basis of their kinematic behavior. The latter was assumed to follow a parametric model of low order. Although widely adopted, such a model is not sufficient to describe non-rigid complex motions such as human motion. Furthermore, depending on the application and the user the goal of the segmentation may vary. It may be the case that motion information alone is insufficient to discriminate between the objects in the scene. Therefore, additional sources of information such as color and/or texture should be used and a model on the color and/or texture properties should be adopted. As far as the order or the generality of the model is concerned, we should note that, in the absence of *a priori* knowledge about the domain, the adoption of a low-order model is more restrictive than in the case of motion. Following the discussion made in the introduction (fig. 1.1) the adoption of a model of a higher order will come at the cost of a higher degree of user interaction.

In this chapter we present a semi-automatic method for labeling image sequences[1]. The goal is the development of a method capable of segmenting complex scenes into objects whose color and motion attributes vary. An outline of the proposed method is illustrated in fig. 5.1. The method operates at three levels. At **Level 1** (pixel level) a feature vector is estimated for each pixel in the current frame. At **Level 2** (segment level) a color segmentation method decomposes the current frame in a number of color segments. Subsequently, we estimate the statistical properties of the color segments under the assumption that the feature vectors at pixels inside the same segment are generated by the same process which is modeled as a multivariate Gaussian. Finally, at **Level 3** (object level) a labeling based on a probabilistic classification of the color segments takes place. The labeling imposes a “common fate” to the pixels belonging to the same color segment. The classification is based on local statistical modeling of the color and motion properties of the objects in the scene. We model the conditional probability of motion and color given the label field, in a window around the center of

---

[1] The basis of this chapter appears in [92]

each segment as a mixture of multivariate Gaussians, each one generated by a different object. By modeling **locally** the distribution and the *a priori* probability of each object we are able to represent objects which exhibit variations in motion, color and spatial characteristics. The classification criterion is the maximization of the joint probability of the label field and the observations (i.e. feature vectors), with respect to the label field. For the maximization a deterministic iterative local search algorithm is employed.

The assumption of **local** homogeneity in color and motion attributes, as opposed to the assumption of **global** homogeneity, allows the modeling of objects with varying characteristics but at the same time it also introduces a degree of uncertainty. The reason is that there is not a unique and robust way of defining an object which attributes are not homogeneous. In order to deal with it, we resort to user interaction. For the first frame of the sequence a description of the local statistical properties of the objects is built based on a label field that is initiated from a user-specified scribble. Then the label field is tracked for the rest of the sequence.

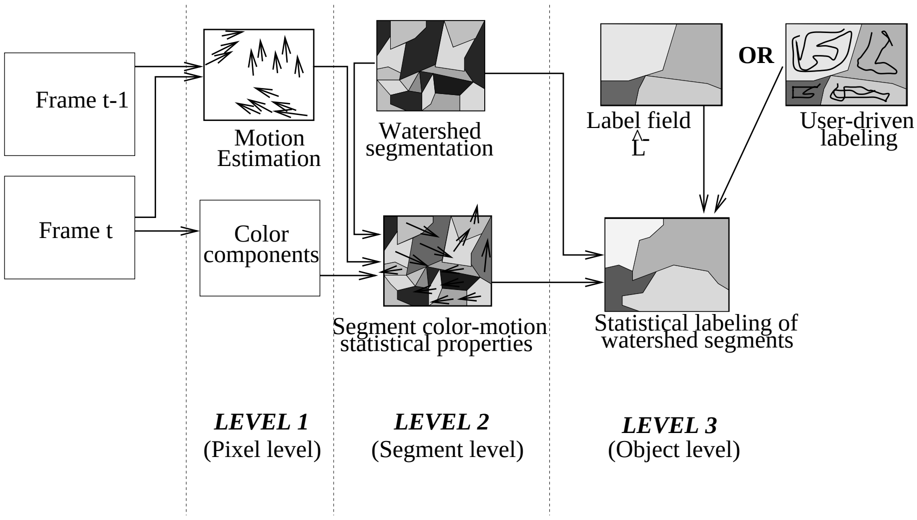


Figure 5.1: Outline of the semi-automatic method

The remainder of the chapter is organized as follows. In section 5.1 we briefly describe the color segmentation method, the user interaction and the motion estimation algorithm. In section 5.2 we describe the local modeling of the feature vectors distribution and in section 5.3 we formulate and solve the labeling as an optimization problem. In section 5.4 the projection procedure is described. In section 5.5 experimental results are presented and in section 5.6 conclusions and discussion follow.

5.1 Feature Extraction and User Interaction

In this section we will discuss the issues related to the first two levels of the proposed method, as well as issues related to the user interaction phase.

At the lower level of the proposed method (**Level 1** in fig. 5.1), for each pixel **i** in the current frame a **feature vector $\mathbf{x_i}$** which characterizes its color and motion properties is defined. The feature vector has five components: one for each of the three dimensions of the chosen color space and two for the horizontal and vertical components of the motion.

As far as the motion estimator is concerned, we should note that the generality of the proposed approach is not limited by the use of a specific motion estimation scheme. However, since we deal with color image sequences we favor motion estimators that incorporate such extra information. In principle, color motion estimators can come as extensions of standard intensity-based motion estimators [44]. While intensity motion estimators attempt to minimize an error measure based on the motion-compensated intensity difference, color motion estimators attempt to minimize an error measure based on the corresponding motion-compensated **color** difference. Consequently, the definition of color difference requires a definition of a distance metric in the adopted color space. In principle, the color distance metric should be dependent on the choice of the color space itself. In the context of our work, we choose the Lu*v* color space and the Euclidean distance. Lu*v* was designed having this distance metric in mind, so that the quantitative measure of differences in color reflect the corresponding perceptual differences.

More specifically, in our motion estimation scheme we aim at a dense motion field, that is, at the estimation of a motion vector $\mathbf{v_i}$ for each pixel **i** in the current frame. We adopt a rectangular support region ($B_\mathbf{i}$) around each pixel in question **i** and aim at the minimization of the motion-compensated mean-square color difference. Formally:

$$\hat{\mathbf{v}}_\mathbf{i} = \arg\min_{\mathbf{v}} \frac{1}{|B_\mathbf{i}|} \sum_{\mathbf{j} \in B_\mathbf{i}} DC(\mathbf{j}, \mathbf{j} - \mathbf{v})^2 \tag{5.1}$$

where $DC(\mathbf{j}, \mathbf{j} - \mathbf{v})$ is the motion-compensated color difference in Lu*v* color space between pixel $\mathbf{j}$ in the current frame and $\mathbf{j} - \mathbf{v}$ in the previous frame. In order to minimize eq. 5.1 we employed two estimation methods. The first is a search scheme commonly employed by block-matching motion estimators ([75], [13]). In the second, the solution is derived by differentiating with respect to the unknown motion vector [60]. The latter scheme has been widely used in the literature in intensity-based motion estimators and is computationally very attractive. In order to deal with motions of large magnitude we incorporated both schemes in a multiscale framework.

At the next level of the proposed method (**Level 2** in fig. 5.1) a statistical representation of the color and motion properties of segments that result from a color segmentation method is built. Like in the previous two chapters, the color segments are the building elements of the objects in the scene, that is, each object is defined as a collection of color segments. For its localization accuracy and its low computational complexity we choose the watershed algorithm to obtain the color segmentation. The latter is applied on the morphological gradient of the color image [100], where the morphological gradient at each point is estimated as the difference between the morphological opening and closing operators, each of which is applied separately to each of the three components of the color frame [45]. Like in the previous two chapters we aim at a conservative color segmentation. As was the case in the motion estimation scheme, we use the Lu*v* color space and the Euclidean distance as a quantitative measure of color differences.

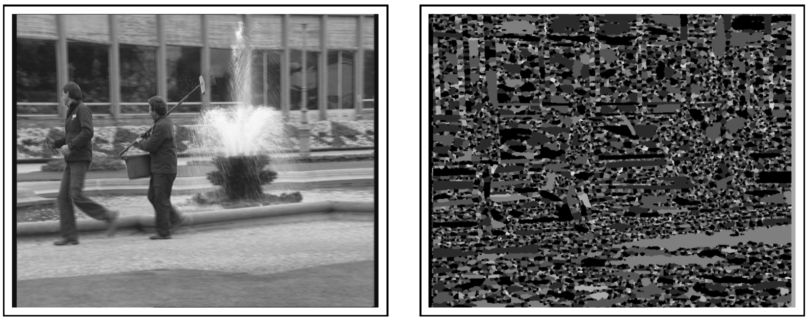


Figure 5.2: An original frame and a watershed color segmentation for the "jardin" sequence.

Once the color segmentation phase is over we build a statistical representation of the color and motion properties of each color segment. Adopting the terminology of the previous two chapters, let us denote with $s \in [1 \ldots K]$ the segments that result from the initial color segmentation phase and with G_s the set of pixels in segment s. We assume that the feature vectors at the points within each segment s are all generated by the same process. Assuming that this process can be approximated within the color segment s with a multivariate Gaussian we estimate its parameters (i.e. mean vector and covariance matrix) ϕ_s with a **Maximum Likelihood** estimator. Assuming that its components are not correlated we consider only the diagonal of the covariance matrix. Let us then define two general operators $\mu_d(.)$ and $\sigma_d(.)$ that return the mean and the deviation of the d-th ($d \in [1, \ldots, 5]$) component of the multivariate Gaussian that receive as argument, respectively. That is, $\mu_d(\phi_s)$ and $\sigma_d(\phi_s)^2$ are the (scalar) mean and the variance of the d-th component of the multivariate Gaussian with parameters ϕ_s.

The assumption that the feature vectors within the same color segment are generated by the same process is violated only in the case that the initial color segmentation violates the object edges. Since the structuring element for the applied morphological operators is very small (3x3) a violation of an object edge is not likely to occur. The latter assumes that a color edge is a necessary condition for an object edge. On the other hand, the assumption that the process that generates each feature vector is a multivariate Gaussian is a design assumption. As such, it can be violated, for example in situations where the motion pattern has large rotational components whose center of rotation is very close to the segment in question. In such cases a parametric model would be more appropriate. However, given that the size of the initial color segments can be rather small in general (e.g. fig. 5.2), a higher-order model would introduce the problem of dimensionality. A low-order model is more robust and less prone to fit to the noise.

At the highest level of the proposed method (**Level 3** in fig. 5.1) each color segment is labeled with an object label in an iterative optimization procedure where the label field and the local statistical properties of the objects in the scene are jointly estimated. Here we will discuss issues related to the initialization of the label field for the first frame of the sequence. For that frame, a user interaction phase is employed. We keep that phase simple, short and intuitive for the user by requiring the drawing of a scribble [22] [119] over the objects in the first frame. We use the user's scribble as markers for a watershed algorithm applied on the color gradient of the first frame. An illustration in the one-dimensional case is depicted in fig. 5.3. Loosely speaking, object edges are declared at the points of highest gradient between the markers. This produces an initial label field which is used for initializing the local model parameters and the local *a priori* probabilities for each object.

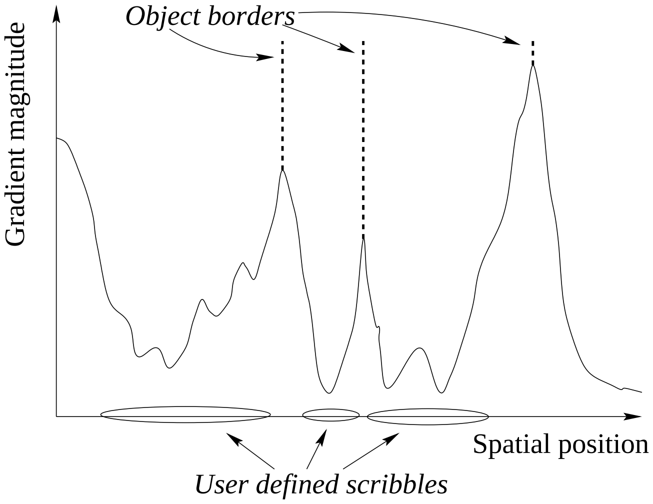


Figure 5.3: User-defined scribbles used as markers by a color watershed algorithm (1-D case)

In comparison to our approach, Chalom [22] also engages a user interaction phase with scribbles. In that work the statistical properties of the objects are estimated only along the scribbles. However, this requires a careful specification of the scribbles to ensure that the feature distribution along them matches the feature distributions of the underlying objects.

5.2 Local Mixture Modeling

At the highest level of the proposed method (**Level 3** in fig. 5.1) the color segments are labeled according to the local statistical color and motion properties of the objects. Statistical representations of the color and motion characteristics of an object have been used in the literature (e.g. [22] [85]) for segmenting and tracking objects in complex scenes. Both in [22] and in [85] a mixture of Gaussians is used to model the color and motion features, and in both of them the labeling is performed per pixel.

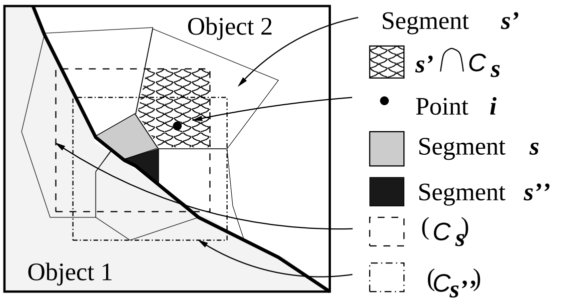


Figure 5.4: Support area for local models

In the framework of our method we model the color and motion characteristics of each object **locally**. More specifically, around each segment s we define a rectangular ($\mathcal{C}_s$) with center the center of mass of segment s (fig. 5.4). This rectangular $\mathcal{C}_s$ is the support area (or neighboring structure) which will be used for building the local distribution of motion and color features for each of the objects present in $\mathcal{C}_s$. The size of the rectangular is adaptive to the segment's size; its dimensions are taken as the multiples of the horizontal and vertical deviations of spatial coordinates of pixels belonging to the segment in question. The multiplicand, denoted with α, is a user-defined parameter that controls the degree of the locality of the model. Assuming that the distribution of the feature vectors for each object n within $\mathcal{C}_s$ can be modeled with a multivariate Gaussian, the total distribution within $\mathcal{C}_s$ is a mixture of Gaussians. Let us denote with S_s the set of segments which intersect with $\mathcal{C}_s$ (i.e. $S_s = \{s' : G_{s'} \cap \mathcal{C}_s \neq \emptyset\}$). Let us also denote with $L = \{l_s : s \in [1 \ldots K]\}$ the label field, where l_s is the label of segment s.

In what follows we will formally define the joint probability of the label field and the feature vectors. An optimization of the joint probability will provide the final label field [108] [5]. Adopting the terminology of the previous chapters, let us assume that there are N objects in the scene, where N is specified by the user in the interaction phase. Let us denote an object label with n, where $n \in [1 \ldots N]$. Let us also denote with $\theta_{sn}(L)$ the parameters of the Gaussian distribution around segment s for the object with label n. Note that in contrast to the previous two chapters θ does **not** denote the parameters of a motion model. Let us also denote with $\pi_{sn}(L)$ the local *a priori* probability of the object n. Note the dependence of $\theta_{sn}(L)$ and $\pi_{sn}(L)$ on the label field. Once the label field is given, $\theta_{sn}(L)$ and $\pi_{sn}(L)$ are estimated with the Maximum Likelihood estimator for each object n that exists in C_s. More specifically, π_{sn} is the percentage of pixels in $\mathcal{C}_s$ with label n. In the same way the mean and the variance components of θ_{sn} are the estimated mean and variance of the data at the points of $\mathcal{C}_s$ with label n. Finally let us define Θ as $\Theta = \{(\theta_{sn}, \pi_{sn}) : s \in [1 \ldots K], n \in [1 \ldots N]\}$. This is the set of the parameters of the local models and serve as the description of the objects' properties. We assume that the *a priori* probability that a pixel $\mathbf{i}$ belongs to an object n is the same for all the pixels $\mathbf{i}'$ in the same color segment s.

Under the above assumptions, the joint probability of the observations ($X = \{\mathbf{x}_\mathbf{i}\}$)

and the label field ($L = \{l_s\}$) is given by:

$$P(X, L) = P(X|L)P(L) = \prod_{\mathbf{i}} p(\mathbf{x_i}|\theta_{sn}(L)) \prod_{\mathbf{i}} \pi_{sn}(L) \tag{5.2}$$

where the pixel **i** belongs to the color segment s and $l_s = n$. The product is taken over all pixels **i** in the current frame.

Once the label field is known the $P(X, L)$ can be estimated in a straightforward way.

The parameter α which controls the degree of locality is the only parameter in our formulation. Concerning this parameter, let us note that in the degenerate case that $\alpha \rightarrow \infty$, then all $\mathcal{C}_s$ contain the whole image and the modeling implies that the feature vectors of each object follow a **unimodal** multivariate Gaussian distribution. In that case, the closer the feature vector to the mean of the object, the higher the probability that the corresponding pixel has the label of the object in question. That model is equivalent to a global, low-order model and is obviously capable of describing with sufficient accuracy a rather limited range of objects, for example only translational motion patterns. On the other hand, as the size of the support area $\mathcal{C}_s$ decreases, the degree of locality of the model increases. In this way we are able to model objects having variations in their color and motion characteristics (e.g. humans). In the way that the support areas $\mathcal{C}_s$ are defined the local modeling implies that the color and motion characteristics are only **locally** homogeneous. Such an assumption is essential for assigning the same label to segments with similar color and motion characteristics. In the extreme case that the support area $\mathcal{C}_s$ is smaller than the color segment itself (i.e. $\alpha \leq 1$), the local model only influences and is influenced by the color segment in question. In that case the label of the color segment does not effect the joint probability of the labels and the observations.

5.3 Maximization of Joint Probability

The labeling criterion is the maximization of eq. 5.2. This is equivalent to the maximization of its logarithm.

$$L(X, L) = \ln(P(X, L)) = \sum_{\mathbf{i}} \ln p(\mathbf{x_i}|\theta_{sn}(L)) + \sum_{\mathbf{i}} \ln(\pi_{sn}(L)) \tag{5.3}$$

We employ an iterative local search algorithm which generates a sequence of label fields L^k (where k denotes the iteration) that increase the log-likelihood function. The optimization procedure also involves the initialization of the label field (L^0). For the first frame this is provided as the result of a watershed segmentation procedure that uses markers extracted from the user-specified scribbles. For each of the remaining frames of the sequence, L^0 is obtained from the estimation of the label field from the previous frame with a projection scheme described in section 5.4. An outline of the whole procedure is depicted in fig. 5.5.

In the iterative local search optimization scheme, given a label field L^k we examine possible perturbations of the label field (L^{k+1}) and calculate the logarithm of the joint

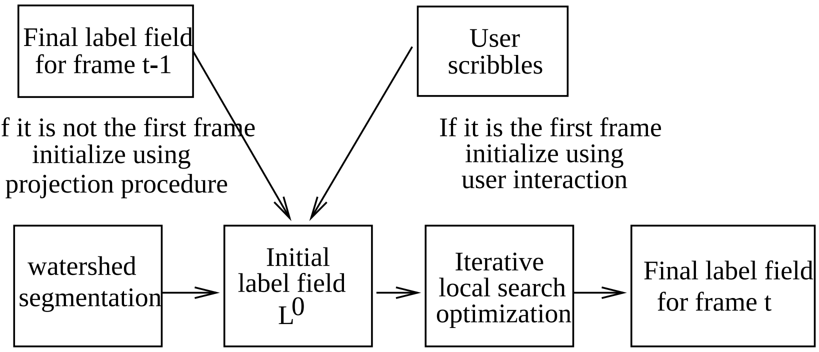

Final label field for frame t-1
User scribbles
If it is not the first frame initialize using projection procedure
If it is the first frame initialize using user interaction
watershed segmentation
Initial label field L^0
Iterative local search optimization
Final label field for frame t

Figure 5.5: Outline of the labeling procedure

probability of the perturbed label field and the observations (eq. 5.3). We accept the perturbations that increase the log-likelihood function. In order to reduce the computational cost we examine perturbations only along the borders between the different objects (as defined by L^k). Furthermore, along the borders of the objects are the locations where misclassifications are to be expected in the initialization of the label field (L^0). For the first frame misclassification occur due to the imprecisions in the user's scribbles. For the rest of the frames misclassifications in the L^0 result due to the shortcomings of the projection scheme, as will be described in section 5.4. The final label field is obtained when no change of the label of a color segment s along the borders, increases the log-likelihood function. Since the log-likelihood function has an upper bound, the stopping criterion is guaranteed to be met. An outline of the iterative local search optimization scheme is depicted in fig. 5.6.

The maximization of $L(X, L)$ with the iterative local search algorithm involves two steps (fig. 5.6). First, by changing the label of a segment s' from n to n', the (θ_{sn}, π_{sn}) and $(\theta_{sn'}, \pi_{sn'})$ for all segments $s : s' \in \mathcal{C}_s$ need to be re-estimated. In the second step, the joint probability of these segments s given the new values of (θ_{sn}, π_{sn}) and $(\theta_{sn'}, \pi_{sn'})$ are estimated. However, these estimations are computationally intensive if they are to be performed at pixel level. We will utilize the statistical representation of each segment to estimate them efficiently. More specifically, we will express them as a function of the segment parameters ϕ_s. To do so we will need some extra notation. Let us denote with $P(s'|C_s)$ the probability of a segment s' given a support area C_s. This can be estimated as the percentage of $\mathcal{C}_s$ that s' occupies (fig. 5.4). That is,

$$P(s'|C_s) = \frac{|G_{s'} \bigcap C_s|}{|C_s|} \tag{5.4}$$

Given the label field and the parameters Θ, let us estimate the log-likelihood func-

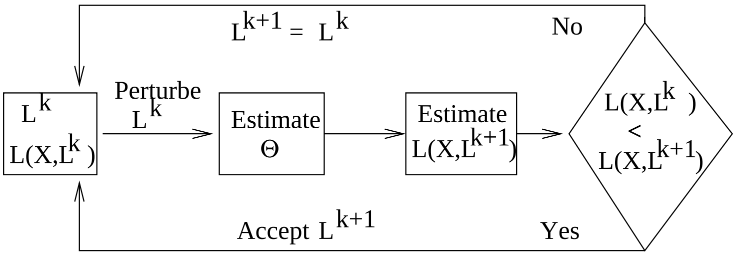


Figure 5.6: Outline of the iterative local search maximization procedure

tion by rearranging eq. 5.3.

$$
\begin{aligned}
L(X,L) &= \sum_{s=1}^{K}\sum_{i\in G_s}\ln p(\mathbf{x_i}|\theta_{sn}(L)) + \sum_{s=1}^{K}\sum_{\mathbf{i}\in G_s}\ln(\pi_{sn}(L)) \qquad (5.5)\\
&= \sum_{s=1}^{K}|G_s|\mathcal{E}_{\mathbf{i}\in G_s}\{\ln p(\mathbf{x_i}|\theta_{sn}(L))\} + \sum_{s=1}^{K}|G_s|\ln(\pi_{sn}(L))
\end{aligned}
$$

For estimating the expected value of $\ln p(\mathbf{x_i}|\theta_{sn}(L))$ we utilize the fact that $\mathbf{x_i}$ are assumed to be generated by a process following the Gaussian distribution parameterized by ϕ_s . By denoting each component of the Gaussian by d and after some algebra we obtain that:

$$
\begin{aligned}
\mathcal{E}_{\mathbf{i}\in G_s}\{\ln p(\mathbf{x_i}|\theta_{sn}(L))\} = & -\sum_{d=1}^{5}\ln(\sqrt{2\pi}\sigma_d(\theta_{sn})) \qquad (5.6)\\
& -\sum_{d=1}^{5}\frac{(\mu_d(\phi_s)-\mu_d(\theta_{sn}))^2+\sigma_d(\phi_s)^2}{2\sigma_d(\theta_{sn})^2}
\end{aligned}
$$

where the dependence of θ_{sn} on L is omitted for notational simplicity.

Let us now estimate the parameters (θ_{sn}, π_{sn}) given the label field L. θ_{sn} is the mean and the variance of the data at points inside $\mathcal{C}_s$ with label n, and π_{sn} is the percentage of points in $\mathcal{C}_s$ with label n. Let us express them as functions of the means and the variances of the segments.

$$
\begin{aligned}
\pi_{sn} &= \sum_{s'\in S_s} P(s'|\mathcal{C}_s)\delta(l_{s'}=n) \qquad (5.7)\\
\mu_d(\theta_{sn}) &= (\pi_{sn})^{-1}\sum_{s'\in S_s} P(s'|\mathcal{C}_s)\mathcal{E}_{\mathbf{i}\in G_{s'}}\{\mathbf{x_i}\delta(l_\mathbf{i}=n)\}\\
&= (\pi_{sn})^{-1}\sum_{s'\in S_s} P(s'|\mathcal{C}_s)\mu_d(\phi_{s'})\delta(l_{s'}=n) \qquad (5.8)\\
\sigma_d(\theta_{sn})^2 &= (\pi_{sn})^{-1}\mathcal{E}_{\mathbf{i}\in\mathcal{C}_s}\left\{(\mathbf{x_i}-\mu_d(\theta_{sn}))^2\delta(l_\mathbf{i}=n)\right\} \qquad (5.9)\\
&= (\pi_{sn})^{-1}\sum_{s'\in S_s}\delta(l_{s'}=n)P(s'|\mathcal{C}_s)\left(\mu_d(\phi_{s'})^2+\sigma_d(\phi_{s'})^2\right)-\mu_d(\theta_{sn})^2
\end{aligned}
$$

Equations 5.7-5.9 provide us in analytical form the parameters of the probability distribution in C_s as a function of the statistical representations $\phi_{s'}$ of the color segments s' that intersect with C_s. In a straightforward manner we can derive from them the difference in these parameters caused by the change of the label of a segment $s' \in \mathcal{C}_s$. More specifically, suppose that the label of a segment s', has changed from n to n'. Then the parameters of the local model for object n around segment s (where $G_{s'} \bigcap C_s \neq \emptyset$) are given by:

$$
\begin{aligned}
\pi_{sn}^{\text{new}} &= \pi_{sn}^{\text{old}} - P(s'|\mathcal{C}_s) && (5.10)\\
\mu_d^{\text{new}}(\theta_{sn}) &= \frac{1}{\pi_{sn}^{\text{new}}}\left[\mu_d^{\text{old}}(\theta_{sn})\pi_{sn}^{\text{old}} - P(s'|\mathcal{C}_s)\mu_d(\phi_{s'})\right] && (5.11)\\
\sigma_d^{\text{new}}(\theta_{sn})^2 &= \frac{\pi_{sn}^{\text{old}}}{\pi_{sn}^{\text{new}}}\left[\sigma_d^{\text{old}}(\theta_{sn})^2 + \mu_d^{\text{old}}(\theta_{sn})^2\right]\\
&- \frac{1}{\pi_{sn}^{\text{new}}}\left[\sigma_d(\phi_{s'})^2 + \mu_d(\phi_{s'})^2\right] P(s'|\mathcal{C}_s) - \mu_d^{\text{new}}(\theta_{sn})^2 && (5.12)
\end{aligned}
$$

Using eq. 5.6 we can then derive the difference in the log-likelihood function.

5.4 Object Projection

One of the most important issues in the segmentation of image sequences is the temporal coherency of the label field. That is, how consistent the label fields are in time. Dealing with this issue involves establishing a link between the label field at the current frame and the label field at the previous frame. This is usually achieved with a motion-based projection in the label field estimated at the previous frame.

In our approach, the temporal consistency affects the label field at the first iteration, that is, the construction of L^0. For each frame of the sequence except the first, each segment s is projected in the previous frame and the parameters θ_{sn} and π_{sn} (for each label n) are estimated locally around the projection area (fig. 5.7). Then, we assign to segment s the label which maximizes the joint probability of the observations in s and its label.

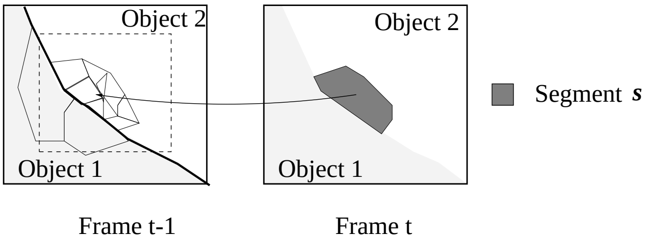


Figure 5.7: Projection of segment s in the label field of the previous frame

More specifically, the window in which the parameters θ_{sn} and π_{sn} are estimated, is constructed around the projection in the previous frame of the center of mass of the color segment s. The dimensions of the support window are determined as multiplicatives of the deviations of coordinates of the pixels belonging to segment s. We should note that it is essential that we do not consider the local model estimated at the previous frame around the color segment to which the segment of mass of s projects. The initial color-based segmentation can vary from frame to frame, thus the dimensions of the support window for the local model estimated in the previous frame may be too small or too large for a correct classification of segment s. For example, in fig. 5.7 segment s is fragmented in smaller segments in the previous frame. The local model, estimated in the previous frame around the segment containing the point of projection, was estimated in a support window which is too small with respect to the dimensions of segment s.

Once the parameters θ_{sn} and π_{sn} are estimated, we assign to s the label which maximizes the joint probability of the observations in s and the label itself. Formally:

$$l_s = \arg\max_n \prod_{\mathbf{i} \in G_s} p(\mathbf{x}_\mathbf{i} | \theta_{sn}) \pi_{sn} \tag{5.13}$$

In the projection procedure the number of the labels is considered known and fixed for the whole sequence. Parts of the objects that are entering the scene are classified successfully if their motion and color characteristics are similar to the ones of the segments that were already visible in the previous frame. However, eq. 5.13 does allow for the classification of segments as parts of an object that just appears in the scene. A new label could be added in case the joint probability of each label n and the observations in s is smaller than a threshold.

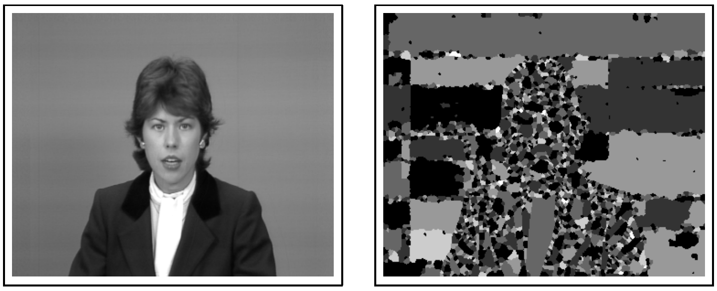
Finally, let us note that in the way that the projection procedure is defined, the effect that the inaccuracies in the projection and the occlusion phenomena have, is limited. That holds because the local models are estimated over larger areas, namely the support windows, which do not have to be localized accurately. It suffices that the support window contains a sufficiently large portion of the object to which segment i belongs. Furthermore, the classification is performed per segment, which makes the procedure robust to a small number of inaccurately estimated features.

5.5 Experimental Results

The method has been tested in a number of image sequences. Here we present results for four of them, namely the “claire” the “mother”, the “jardin” and the “coast guard” image sequences.

For the “claire” sequence we present results for two settings. In the first we decompose the scene into two objects, namely the woman and the background. In the second, the face of the woman is marked at the first frame as a third object. In fig. 5.8 we present the first frame of the sequence and the watershed color segmentation and fig. 5.9 depicts the user scribbles for both settings as well as the initial label field L^0 for the three objects setting are depicted. It is clear that the conservative watershed segmentation provides good edge localization and that object edges are not violated. Furthermore, the user scribbles need not to be very accurate in the case that we aim at a segmentation in two objects, since the color structure in the background is rather poor.

In the case that a separation of the face from the body and the hair is desired, a more careful initialization is required.

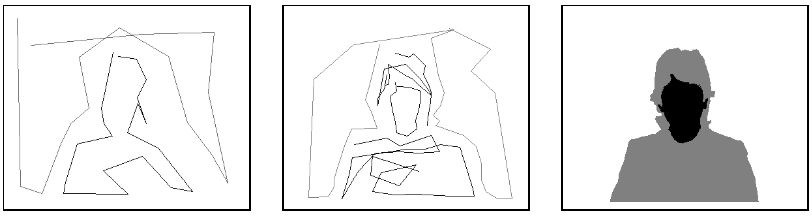


(a) Original frame

(b) Color watershed segmentation

Figure 5.8: Original frame and color segmentation of “claire” image sequence

In fig. 5.13 we present the labeling results for 75 frames with a step of 15 frames. In the first and the second column the label fields for each of the two cases are presented. In the third column the borders of the objects are superposed on the corresponding original frames to demonstrate the localization accuracy of the method. In both cases the label fields were very stable in time and the borders of the objects were well localized.



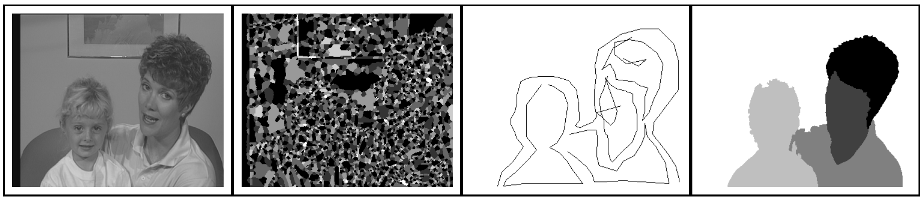
(a) Markers for two objects

(b) Markers for three objects

(c) Label field L^0 for three objects

Figure 5.9: User-specified markers and label field L^0 for “claire” sequence

For the “mother” sequence, a segmentation into five objects was attempted, namely the background, the child, the face, the hair and the body of the woman. fig. 5.10 depicts the first frame of the sequence, the watershed segmentation, the user-defined scribble and the initial label field L^0. In fig. 5.14 we present the segmentation results for 75 frames with a step of 15 frames. The first column contains the label fields and the second column the borders of each object are superposed on the original frames. The label field is quite accurate and the borders well localized. However, a number of misclassifications is also visible. In the frames between frame 45 and 70 (three last rows of fig. 5.14) sometimes a part of the picture in the background has merged with the hair of the woman. This artifact is due to the absence of a significant color edge, as a result of which the initial color segmentation produces a single segment which contains both part of the hair and part of the picture in the background. Since the misclassified area has a thin and elongated shape, simple morphological operations on the final label field would remove the problem in this specific case [101]. A more general remedy would be to introduce temporal constraints in the initial color segmentation or in the final label field. However, such solutions require accurate projections of the borders, which can be problematic either due to inaccuracies in the motion estimation or due to occlusions. For the same sequence, misclassifications occur when the hand of the woman is introduced to the scene due to the absence of reliable temporal constraints.



(a) Original frame

(b) Markers for five objects

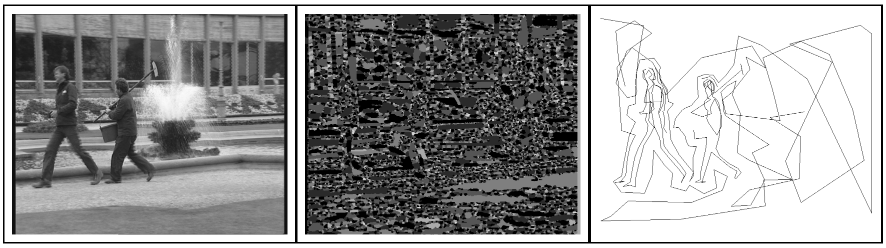
(c) Markers for five objects

(d) Label Field L^0 for five objects

Figure 5.10: “mother” sequence: results for the first frame

In the “jardin” image sequence we are presented with a complex scene in terms of both motion and color characteristics. In terms of motion the scene contains three distinct motion patterns: The apparent background motion due to the camera motion, the quite arbitrary motion of the water of the fountain and the non-rigid human motion of the two walking men. In terms of color, the characteristics of both the background as well as the two men exhibit variations. For this sequence, we attempt a segmentation into three objects, namely the two men and the background. The first frame of the sequence, the corresponding watershed color segmentation and the user markers are depicted in fig. 5.11. Since the scene is rich in color structure the conservative initial color segmentation method results in an oversegmentation of the scene. On the other hand, the edge localization is very good. The scribbles defined by the user interaction are depicted in fig. 5.11(c).

The results of the labeling for 15 frames of the sequence are depicted in fig. 5.15. In the first column we present the label fields and in the second column the borders of each object are superposed on the corresponding original frames. The proposed method exhibits good localization properties and is rather stable in time. Misclassifications that occur in disclosured areas (e.g. behind the two men) do not propagate to the subsequent frames. The main misclassifications occur in thin, elongated areas, such as the broom that is carried by the second man (note that the user’s scribble marks



(a) Original frame

(b) Color watershed segmentation

(c) Object markers

Figure 5.11: Original frame, color segmentation and user's scribbles for the "jardin" image sequence

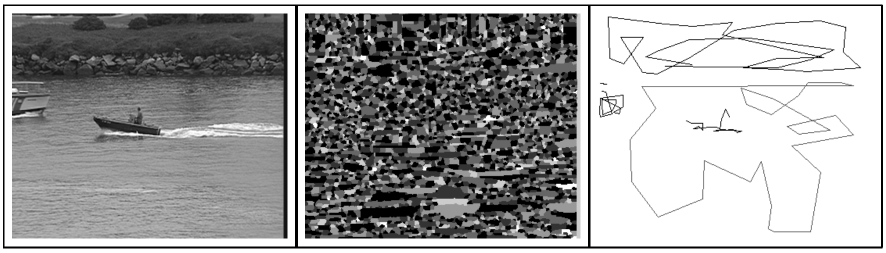
the broom as part of the second man) and in areas that just appear in the scene. The dimensions of the support areas that have been used are 3.5 times as large as those of the corresponding color segments. As a consequence in areas like the broom, the homogeneity of the label field (implicitly imposed by the π_{sn}) overcomes the color and motion evidence (the latter not being very reliable). The spatial homogeneity is also the reason that occasionally areas like the head and the leg of the first man merge with the background. On the other hand, much smaller support windows produce noisy label fields with isolated segments.

Finally, we present results for the image sequence "coast guard". The first original frame, the corresponding watershed color segmentation and the user's scribbles are depicted in fig. 5.12. In fig. 5.16 we present the label fields and the corresponding superpositions of the object borders on the original frames for 45 frames of the sequence with a step of 9 frames. The proposed method exhibits very good temporal stability and localization properties.

5.6 Conclusions

In this chapter we presented a semi-automatic method for labeling image sequences based on a local model-based statistical classification algorithm. An initial color segmentation scheme partitions each frame in a number of segments which are subsequently labeled on the basis of their color and motion statistics. The labeling is expressed as an optimization problem, where the criterion is the maximization of the joint probability of the labels and the color and motion distribution within each object. Limited user interaction is required for the first frame of the sequence.

Experimental results have been presented for four image sequences containing objects whose motion and color attributes show variation. We have tested our method in complex scenes with moving cameras and highly cluttered background. We have presented results for rigid as well as non-rigid human motion without *a priori* knowledge



(a) Original frame

(b) Color watershed segmentation

(c) Object markers

Figure 5.12: Original frame, color segmentation and user's scribbles for the "coast guard" image sequence

of the contents of the scene. The label fields that were obtained were temporally stable and localized quite well.

The principle of the proposed method is quite simple and a number of extensions could be incorporated to increase its performance and to deal with its weaknesses. In our opinion future research should aim at two main directions. The first one involves the treatment of objects entering the scene (e.g. the hand of the woman in fig. 5.14), which in the current framework are assumed to belong to one of the objects already present in the scene. Even in the cases that such an assumption holds true a correct classification is problematic since the entering object does not necessarily possess color and/or motion characteristics similar to those of the object to which it belongs. However, identification of such areas and the assignment to them of a new label might be an easier task. A new label can be added if the joint probability of each label and the observations in the segment in question (eq. 5.13) is smaller than a threshold.

The second direction involves a model for the temporal evolution of the statistical representations of the objects' properties or a temporal link between subsequent label fields. Both model and link aim at higher temporal stability at the cost of a lower degree of adaption to the data. The first could be achieved in terms of Kalman filtering [20] of the parameters of the Gaussians; the second by introducing temporal dependencies in the *a priori* probabilities of each label π_{sn}, that is, by defining them as conditionally dependent on the label field of the previous frame.

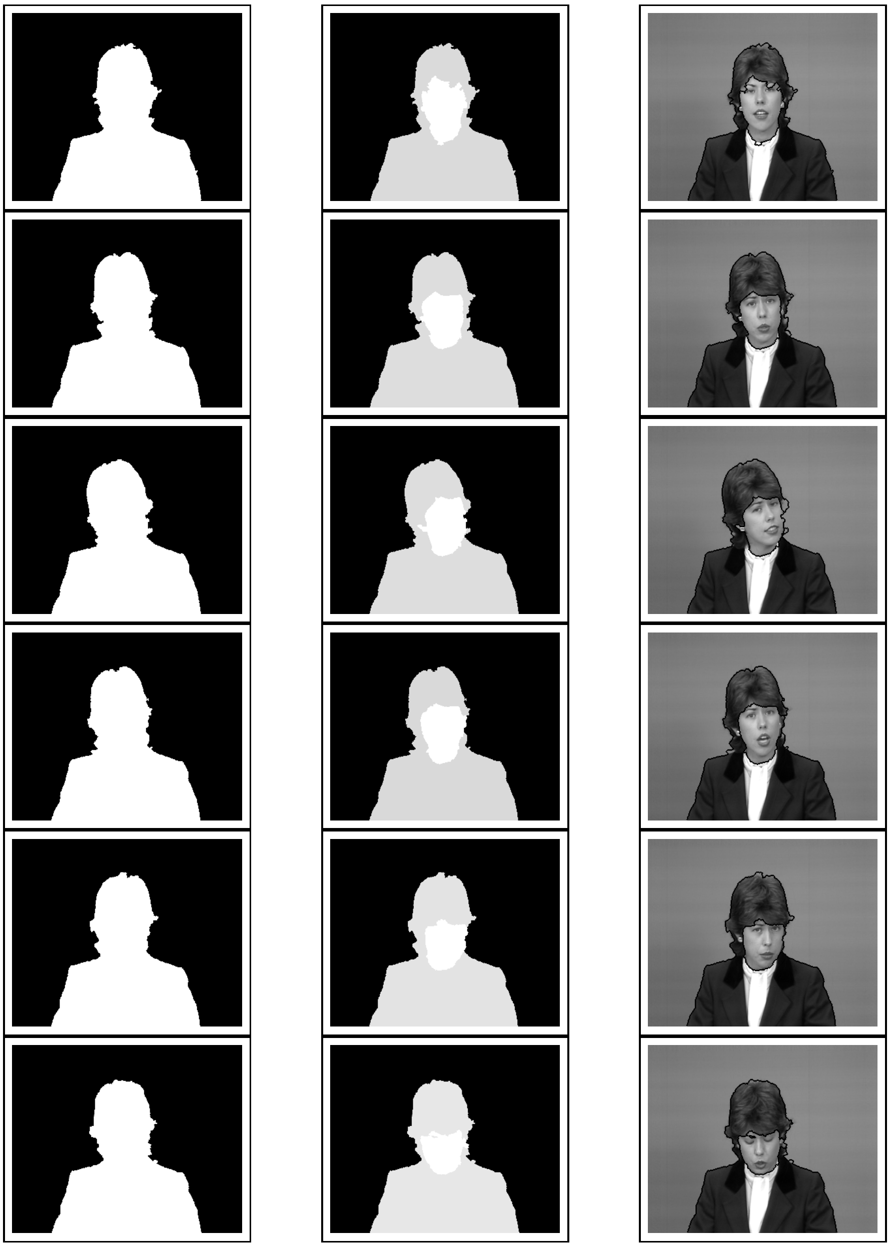


Figure 5.13: Results for the “claire” sequence. First Column: The label field for segmentation in two objects. Second Column: Label field when the face is considered a separate object. Third Column: Superposition of the contours of the face and the body on the original frames of the sequence. Results are shown for frames 3, 15, 30, 45, 60 and 75

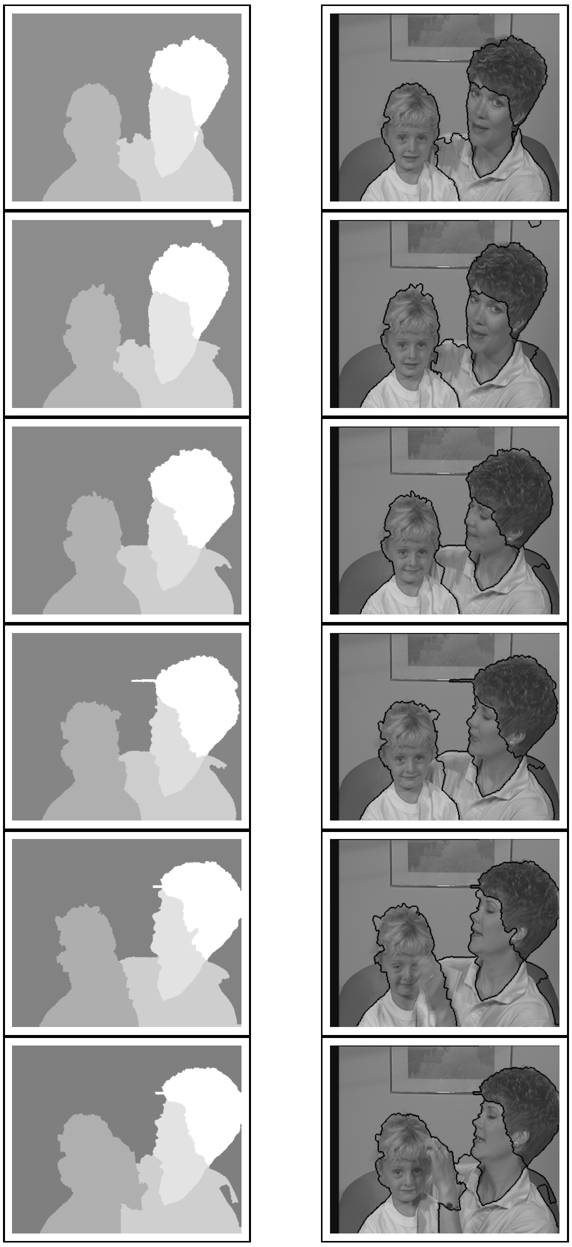


Figure 5.14: Results for the “mother” sequence. First Column: The label field for segmentation in four objects. Second Column: Superposition of the contours of the objects on the original frames of the sequence. Results are shown for frames 1, 15, 30, 45, 60 and 75

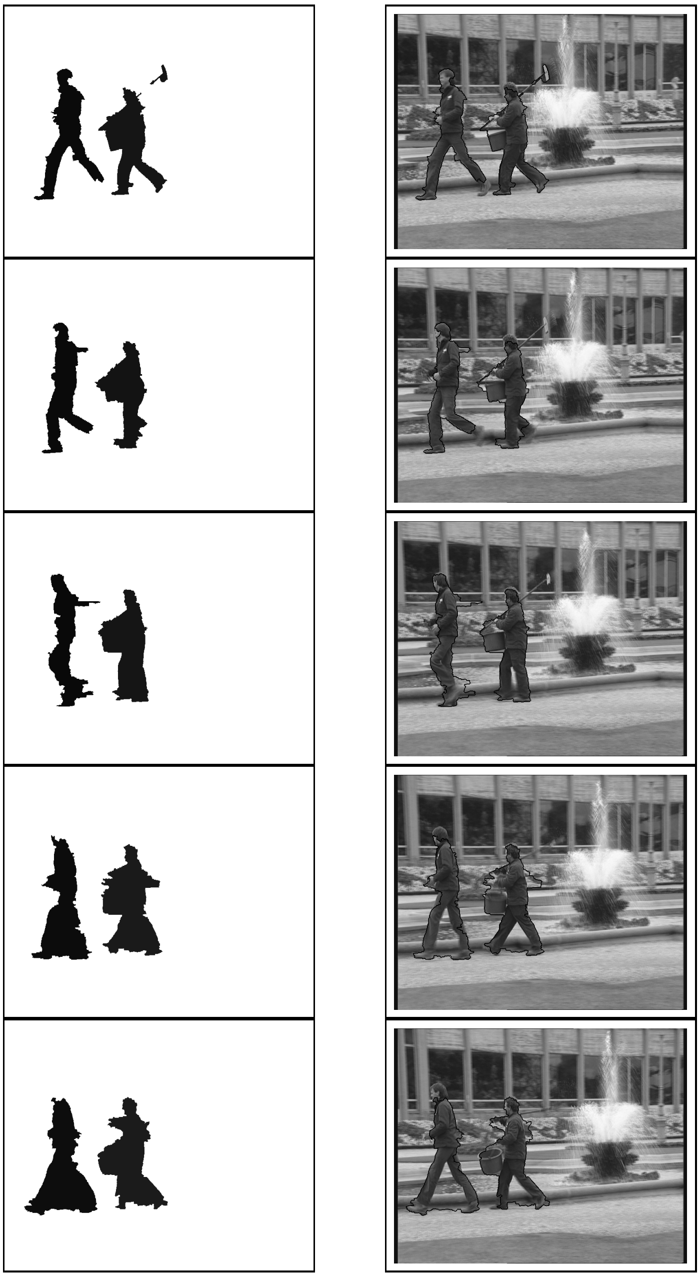


Figure 5.15: Results for the “jardin” sequence. First Column: The label field for segmentation in three objects. Second Column: Superposition of the contours of the objects on the original frames of the sequence. Results are shown for frames 1 to 15 with a step of 4 frames. ($\alpha = 3.5$)

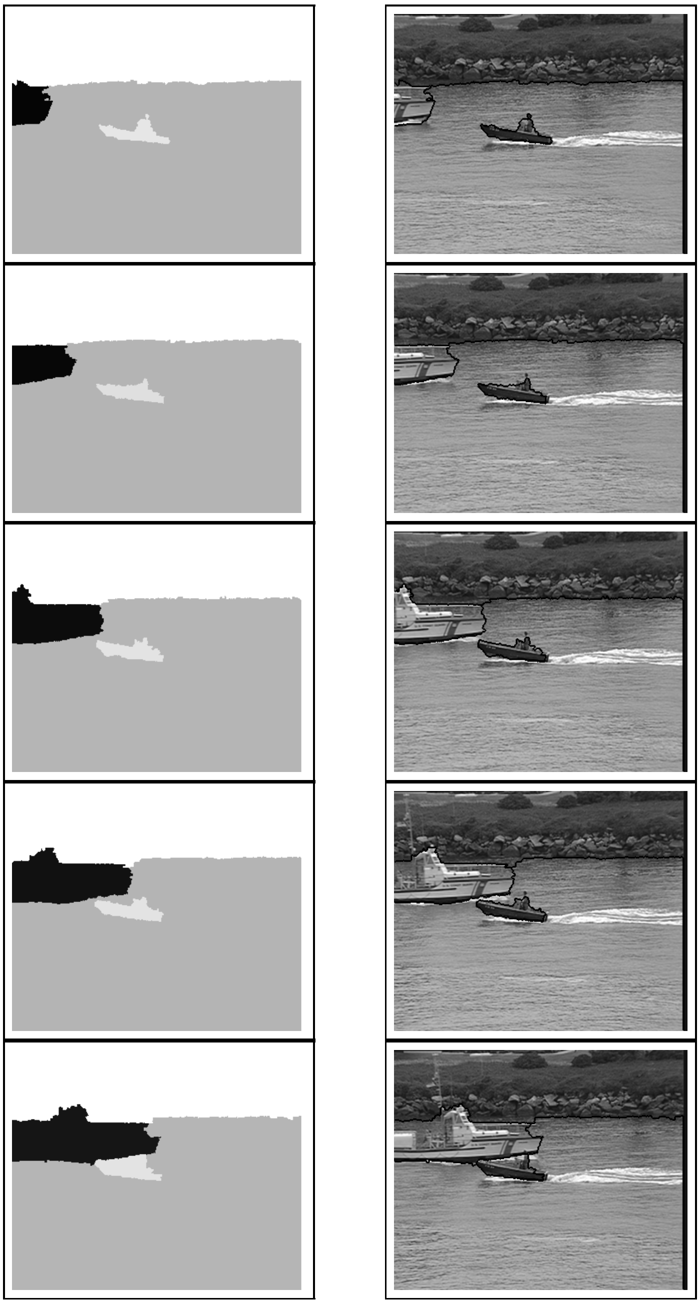


Figure 5.16: Results for the “coast guard” sequence. First Column: The label field for segmentation in four objects. Second Column: Superposition of the contours of the objects on the original frames of the sequence. Results are shown for frames 10, 19, 28, 37 and 46 ($\alpha = 3.5$)

Chapter 6

Discussion

In the previous three chapters we proposed three methods for object-based segmentation of image sequences, while in the first two chapters we gave an introduction and a review of the related work which positioned our work in the field. In this chapter we attempt an *a posteriori* discussion/evaluation of the design decisions that we made. The purpose of this chapter is two fold: On the one hand we outline our contributions. On the other hand, such an evaluation serves as a starting point for a discussion for future research. Such a discussion extents beyond improvements of the proposed methods. It rather expresses the author's personal view on some of the major issues in the field. As such, it is not intended as an exhaustive review and evaluation of the different approaches in the field.

The chapter is organized in a hierarchical way in which we first discuss common design decisions in the development of the proposed methods and later we discern between them based on their individual characteristics. The organization of the discussion is depicted in fig. 6.1, where the nodes of the tree represent the discussion items. At the higher level of the hierarchy we discuss issues related to our proposed paradigm: "fine initial intensity segmentation" - "labeling based on models that consider the intensity segments as primary elements". Subsequently, we discern between three main directions depending on the level and type of user interaction that they require. Within the framework of the first direction we discuss about fully automatic methods. Within the framework of the second direction we discuss on issues raised when the goal is medium-level user interaction. The motion-based segmentation methods that we described in chapter 3 and chapter 4 follow that direction. Subsequently, we discern between them according to the paradigm they adopt: while the method of chapter 3 separates the motion hypothesis extraction phase from the labeling phase, the method of chapter 4 estimates jointly the motion hypotheses and the label field. Finally, within the framework of the third direction we discuss issues raised for methods that require higher level of user interaction in order to deal with more complex scenes. Since the method of chapter 5 follows that direction, we build the discussion around its advantages and limitations.

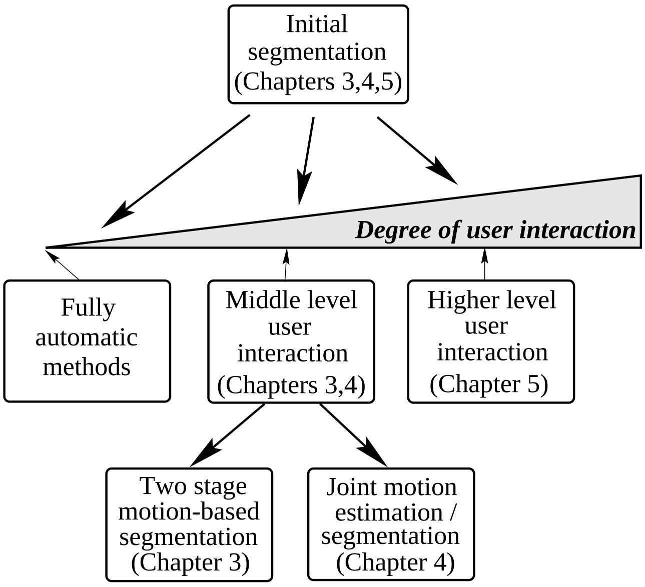

Initial segmentation (Chapters 3,4,5)
Degree of user interaction
Fully automatic methods
Middle level user interaction (Chapters 3,4)
Higher level user interaction (Chapter 5)
Two stage motion-based segmentation (Chapter 3)
Joint motion estimation / segmentation (Chapter 4)

Figure 6.1: Organization of chapter 6. The discussion traverses the tree in a depth-first way where the nodes of the tree represent the discussion items

6.1 Intensity segments as primary elements

In this dissertation we have proposed methods that utilize an independent “fine” initial intensity (color) segmentation. The resulting intensity (color) segments are used as primary elements for the subsequent processing. The proposed methods come to cover the gap between pixel-based approaches and segment-based approaches where the initial segments are relatively large. Let us discuss on the merits and the disadvantages of each.

At the one end of the spectrum are the pixel-based approaches. In comparison to the segment-based approaches which lie at the other end of the spectrum, they are characterized by:

(a) High dimensionality of the solution space.

(b) High degree of ambiguity about the labeling of a single pixel due to uncertainties and/or inaccuracies in determination of features such as motion-related ones. This leads to higher dependence on spatial and/or temporal constraints.

These characteristics usually lead to objective functions with a lot of local minima which are optimized via complicated optimization schemes.

At the other end of the spectrum, larger segments[1] can reduce the dimensionality of the solution space, disambiguate the estimation of properties such as motion and

---

[1] To be more precise, the size of the segments is not always the issue in such methods. What is relevant is that the initial segmentation provides segments for which the properties on which the labeling is performed,

need to employ optimization schemes which are, by far, less computational expensive. In comparison to pixel-based approaches they bring only advantages under a single critical assumption: that the initial segments do not violate object edges.

In our view, it is difficult to guarantee that an intensity segmentation scheme, which is typically employed in order to obtain such an initial segmentation, will respect such an assumption for a large number of image sequences. Complex scenes which include objects with high within-object intensity / color variation are difficult to be automatically segmented in a small number of intensity segments. Even the conservative approach that we employed failed in a number of situations.

In this dissertation we proposed methods that utilize an initial intensity (color) segmentation that in terms of granularity lie in the middle/high end of the spectrum. In order to deal with the ambiguity that, possibly small, initial segments may introduce we introduce constraints in terms of models that consider the initial segments as primary elements. More specifically, in chapters 3 and 4 we proposed models that consider intra-segment dependencies on the label field, while in chapter 5 we proposed the use of local models whose region of support extents beyond the borders of the color segments.

The intensity (color) segmentation that we proposed introduces information about an intensity (color) edge in an early stage. In general, the assumption that an object edge requires the presence of an intensity edge can be violated and we will comment on that latter in this section. On the other hand, such an assumption is valid for the large majority of video material and to our opinion, is useful to be employed. The conservative approach that we applied for the initial intensity segmentation was justified by good localization of the initial intensity (color) segments. Furthermore, the use of intensity segments as primary elements increased the robustness and decreased the computational complexity in comparison to the corresponding pixel-based approaches. In chapter 3 we have shown comparative results that support our claim as far as robustness and computational complexity are concerned.

The use of an initial segmentation poses problems to the extent that object borders are violated. Such a situation might occur when presented with image sequences with objects with high degree of texture. An extreme example is motion of random dot patterns which a human observer can effortlessly segment. A possible remedy for such situations, and a possible extension of our approach, is to consider the potential break up of an intensity segment with criteria such as the degree of conformity to the motion model.

Finally, let us note that the principle of the approach that we propose is quite simple and can be possibly applied in other similar problems not necessarily from the same domain. That holds mainly for the methods proposed in chapters 3 and 4 since the method of chapter 5 is more oriented to the the domain of video segmentation. If the motion-based segmentation is seen as a clustering problem, we propose a two-stage approach. At the first stage we form small initial clusters based on a conservative local grouping. At the second stage we apply a clustering method with primary elements the initial groups. Since the dimensionality of the problem is reduced in the second stage,

---

are reliable enough. Since, usually, that includes the reliable estimation of a parametric motion model, the segments need to be rather large. Of course, depending on the method, different degrees of dependence on the accuracy of the estimated properties are reported more powerful clustering schemes can be employed. Since the initial grouping might be erroneous it would be wise that such schemes address robustness issues. The advantages of such an approach was clearly illustrated in the motion hypotheses extraction phase of chapter 3. However, since we do not have experimental results on data from another domain, such a discussion should be seen in the context of possible extensions. Furthermore, we should note that in our domain, the feature that has been used for the initial conservative grouping has very good localization properties.

The methods proposed in chapters 3 and 4 attempt an object-based segmentation adopting **global parametric models** of low order of the objects' motion characteristics. On the other hand, the method of chapter 5 bases the labeling on the **local statistics** of color and motion. This change in the complexity of the assumed model comes as a trade-off between automation and the complexity of the object's properties. In what follows we will attempt a discussion on issues raised for three main categories of approaches: fully automatic methods, methods that require low-level user interaction and methods that require higher-level user-interaction.

6.2 Fully automatic methods

Fully automated methods that can be applied in all image sequences are of course a chimera. The main reason being the ambiguity at the semantical level of the purpose of the segmentation. Even for the same image sequence the desired results may vary, depending on the application. The development of a system which can detect the context and subsequently deduce the purpose of the segmentation and the semantics of the scene is still far from realization.

On the other hand, fully automatic methods have been reported to work well in the case that the domain on which they are applied is restricted. The latter means that either the contents of the scene are known in advance (e.g. faces), or rather strict assumptions are made for the patterns of the properties of the objects (e.g. static background in surveillance applications). Although the domain is restricted, the applications that arise may be of great importance. Furthermore, the constraints that the domain knowledge introduces usually have great positive impact on the robustness and speed of the resulting schemes. These can be of critical importance for some applications.

Apart from the difficulties and the challenges that are inherent to each domain, an important issue that arises is how domain knowledge is introduced in the modeling. This includes the construction of domain-specific models and addresses the user's role in the design procedure. Advances in that area can be of importance both for scientific development as well as for an end-user since they will allow an easier and faster change between domains. The employment of rather general models whose parameters and components are determined by training in examples from the specific domain are, according to the author, steps in the right direction. However, to the best of our knowledge, issues related to the user interaction to the modeling procedure are not yet sufficiently explored.

6.3 Medium-level user interaction

With the term medium-level user interaction we mean the tuning by the user of a small number of parameters. That is in contrast to higher level interaction that offers implicit information about the contents of the scene by specifying the localization of the objects of interest. Such parameters can be the number of objects or a parameter related to the degree of within-object homogeneity in the properties on which the labeling is performed.

As far as the properties on which the grouping / labeling is based, according to the author, the best, so far, proposed features are related to the temporal behavior of the objects. Such are motion-related features, and temporal constraints on the localization of the objects. Other properties, such as depth are also important, but for monocular image sequences they are usually derived from motion-based features.

Motion related properties were also utilized in the methods developed in chapters 3 and 4. Let us discuss on some of the common design choices.

Affine motion models Affine motion models come as a good compromise between complexity and robustness and have been widely utilized in the context of motion-based segmentation. However, the rigidity assumption that they introduce imposes a limitation which in some cases can be too restrictive for object-based segmentation. Such an assumption can be overcome, in principle, by methods that assume that the motion field is smooth within the objects and detect object borders at motion discontinuities. Let us briefly outline the advantages of each approach as well as the challenges that it should face.

On the one hand, global parametric models of low order impose rigidity assumptions. On the other hand, the estimation of their parameters is object-based, that is, it is based on large regions formed by collections of primary elements (pixels or segments). The use of collective constraints over the large regions robustifies the estimation of the parameters of the models significantly. At the same time, the use of a global model allows the utilization of the motion-compensated intensity differences for labeling. The latter, does not introduces the inaccuracies that an independently estimated feature, such as independently estimated motion, does. In both of the methods of chapter 3 and chapter 4 we were able to estimate the motion parameters with a degree of accuracy that allowed a good labeling based on the motion-compensated intensity differences.

The main, yet very important, advantage of using motion-smoothness constraints is that such a model covers objects with non-rigid motion patterns. On the other hand, such approaches require high reliability of the motion properties that are estimated for the primary elements (pixels or segments). It is difficult to estimate accurately such properties at the areas that accuracy is needed the most: at the borders between the different objects. Such inaccuracies are mainly due to motion-generated occlusions and their extent depends also on the size of the primary elements. Since the primary elements need to be rather small, to our opinion, such methods should face the challenge of the joint estimation of the motion properties and the label of a segment. Spatial constraints should be added for regularization purposes which perform a sort of smoothing and at the same time respect discontinuities: a task which by no means is trivial. This is an interesting direction of future research on which we will elaborate further in section 6.3.2.

Markov Random Fields on the intensity segments In both of the methods of chapters 3 and chapter 4 we modeled the label field as a Markov Random Field where the segments that resulted from an initial intensity segmentation were used as the sites in the formulation. Such modeling has been widely used in pixel-based approaches to express spatial and/or temporal constraints. In our case, it helped to produce clean label fields without isolated intensity segments for both of the proposed methods. Such a model expresses the local interactions between the sites (intensity segments) in a way that results in a well-defined global optimization criterion. Furthermore, the formulation allows the application of optimization methods that allow to reverse the assignment of a label to a specific site (intensity segment in our case). This property is particularly important when the segment's properties need to be estimated jointly with the label field (i.e. for the method of chapter 4).

In both of the methods we proposed that the interaction between intensity segments is proportional to the length of their common border. In general, such a decision puts the emphasis on the spatial constraints for smaller intensity segments, while for larger segments the emphasis is put on the temporal constraints. It would be interesting to investigate and possibly adjust such interaction depending also on the shape of the intensity segments.

Alternative approaches for the application of spatial constraints include region merging methods and level-set approaches. However, in both the labeling of an intensity segment is irreversible. On the other hand level-set approaches offer significant advantages in terms of computational complexity and their application with intensity segments as primary elements could be an interesting direction of research. Their application once the motion hypotheses are reliably estimated seems straightforward. On the other hand, a joint motion estimation / segmentation scheme does not seem easily feasible.

A three-frame approach on motion-compensated intensity differences In both of the proposed methods the motion compensated intensity differences have been used as evidence in the labeling phase. In chapters 3 and 4 we have clearly illustrated the advantages of using such evidence for labeling instead of an independently estimated feature such as motion. The main reason is the inevitable inaccuracies that the estimation of the latter introduces, especially in the areas near the borders of the objects. On the other hand, we should note that using as evidence the motion-compensated intensity differences introduces a larger number of local minima in the objective function which can be problematic in that case that the motion hypotheses are far from the correct ones. The latter is evident from the higher degree of dependence on the initializations of the method in chapter 4 in comparison to the method in chapter 3. In order to overcome the latter, multiscaling approaches should be adopted.

In both of the methods we proposed a three-frame approach where evidence for the temporal behavior of the objects were sought either in forward or in backward direction. This offered a significant improvement in labeling accuracy. Such an approach is particularly beneficial in the presence of motion large in magnitude as was illustrated both in chapter 3 and chapter 4.

The methods of chapter 3 and 4 differ on the adopted paradigm for the motion hypotheses extraction and motion-based labeling. The first one adopts a two-stage approach in which the motion hypotheses extraction and the labeling are performed independently from each other. The second method incorporates all of the constraints in a single framework and attempts the joint estimation of the motion hypotheses and the label field. While the method of chapter 3 estimates the motion hypotheses from an independently estimated motion field, the second seeks evidence directly in the image intensities.

6.3.1 Two-stage motion hypotheses extraction and motion-based labeling

The method proposed in chapter 3 separates the motion hypotheses extraction phase from the labeling phase. In contrast to most of the methods that utilize a dense motion field for the motion hypotheses extraction, we propose the use of the motion-compensated intensity differences as evidence of the conformity of the kinematic behavior of an region with a motion hypotheses. As we have already discussed in the previous section, we have clearly shown that once the motion hypotheses are well estimated, the motion-compensated intensity differences offer far better evidence for the labeling than the motion field itself.

For the motion hypotheses extraction phase a clustering algorithm has been proposed. The challenge that had to be faced is that of the estimation of the parameters of a number of models in presence of inaccurate data. In order to deal with the inaccuracies of the motion estimation field we have proposed confidence measures derived from an analysis that expresses the utilized motion estimator in a probabilistic framework. We have clearly demonstrated the relevance of the derived confidence measures. For the clustering we have proposed an extension of the C-Means algorithm that incorporates the use of the derived confidence measures, robust statistics and the initial intensity segmentation. We have clearly illustrated the benefits of the proposed scheme in terms of computational complexity, accuracy and robustness in presence of inaccurate motion fields.

In terms of the computational complexity we have clearly shown the advantages of using intensity segments as primary elements instead of pixels. We have derived intermediate measures that allow a computational complexity proportional to the number of intensity segments at each iteration. In contrast to other methods in the literature this does not involve the estimation of affine parameters for each intensity segment and clustering in the parameter space which is sensitive to the parametric representation of each intensity segment.

Further research in this direction should address the dependence of the clustering on the initializations. Stochastic optimization procedures might provide an answer but their incorporation in the optimization procedure should respect the differences in the statistics of pixels in comparison to statistics of intensity segments. To be more specific there is a difference in the landscape of the energy function in the solution space specified by the motion parameters and the label field. In the direction of the label field, the energy function has a more step-wise behavior since the change of the label of an intensity segment implies a higher difference in the energy function in comparison to the corresponding difference that the change of the label of a pixel would introduce. Moreover, such a difference depends (implicitly) on the size of the intensity segment.

Another interesting direction of future research would be towards the automatic determination of the parameter z_c which is related to the strength of the spatial interaction between neighboring segments. Although the majority of the approaches that use a Markov Random Field modeling do not address the issue of its automatic determination, conceptually, it seems rather redundant once the number of the objects is specified. Furthermore, since the Maximum Likelihood labeling already provides a reasonably good initial label field, its automatic determination with cross validation techniques seems feasible.

Finally we should note that the applicability of methods that assume smoothness on the motion field can be successfully applied to an independently estimated motion field to the degree that the latter is accurately estimated. In the case of a simple (e.g. block-based) motion estimation method it is preferable, to our opinion, that such a method is applied to obtain only an initialization of the label field. In a subsequent stage, motion properties near the borders should be simultaneously estimated with the label field at the corresponding areas.

6.3.2 Joint motion estimation and segmentation

The method of chapter 4 proposed the incorporation of the spatial and temporal constraints on the label field and on the image intensities in a single framework. The problem is expressed as an optimization problem in terms of the *a posteriori* probability of the label field. Each of the constraints is expressed in terms of an assumption about the corresponding probability distribution. The advantage of such an approach is that it separates the modeling from the optimization procedure and that it allows the addition/modification of the constraints in terms of the assumptions for the underlying distributions. The incorporation of all the constraints in a single framework allows the joint estimation of the motion parameters and of the label field, thus exploiting the interdependencies between them. In order to jointly estimate the motion parameters and the label field such a unified framework with a global objective criterion seems to be a good modeling choice.

The spatial and temporal constraints were expressed by modeling the underlying distributions as Gibbs distributions. That choice allowed us to exploit the extensive work that has been conducted in the field of MAP-MRF modeling (Maximization of the *a posteriori* probability under Markov Random Field modeling). This modeling seems particularly suited for the joint estimation of the label field and the motion parameters. Furthermore, it allowed relatively easily the incorporation of the three-frame extension that we have proposed and the introduction of the **direction** field in the optimization procedure.

We have opted for deterministic methods and for hard decisions in the optimization scheme. This in principle implies higher dependence on the initializations at the gain of lower computational cost. As far as extensions to the optimization procedure are concerned, there are three directions that we would consider worth of further investigation.

Multiple scales Multiscale extensions aim to reduce the sensitivity of deterministic methods on the presence of local minima in the objective function and to the reduction of the computational cost. Such extension which is widely used for pixel-based approaches is not straightforward when intensity segments are used as primary elements. The reason is that it is not trivial to define what the intensity segmentation should be in a higher scale. For this reason, we opt for an extension that performs the labeling on the original intensity segmentation but derives the constraints from different scales of the original image sequence (fig. 6.2).

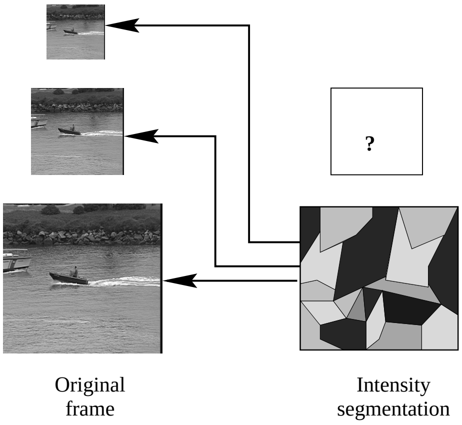


Figure 6.2: Extension to multiple scales. Constraints on multiple scales of the original frames are applied for the labeling of the intensity segments.

Stochastic optimization Such methods (e.g. simulated annealing) overcome the presence of local minima in the objective function by accepting with certain probability changes in the parameter space that increase the objective function. Such extensions could be applied also in the method that we developed as long as they take into consideration that the landscape of the objective function has higher and more irregular discontinuities in correspondence to pixel-based methods.

Soft decisions The incorporation of soft decisions in the labeling is practically straightforward as we show in appendix C.

Possible extensions of the proposed method include the automatic determination of the parameters z_c and z_t, the anticipation of objects entering the scene and the use of a parameter related to the motion homogeneity as an alternative to determining the number of objects. However, the most interesting direction, to our opinion, is the relaxation of the rigidity assumption and the adoption of smoothness constraints on the motion field. In that direction we have already developed a motion estimation method that imposes anisotropic smoothness constraints with intensity segments as its primary elements and we have obtained very encouraging results. The challenges in this direction lie in the way that the smoothness constraints on the motion field are applied when intensity segments are considered as primary elements. This, includes the modeling of the kinematic behavior of each intensity segment.

6.3.3 Open issues in motion-based segmentation

There are several issues in the context of motion-based segmentation that are not addressed by the proposed methods and some that, to our opinion, are not sufficiently addressed in the related literature either. A non exhaustive list includes:

Temporal constraints in more than two frames Our methods, as well as the majority of the methods in the literature, express the temporal constraints by considering fixed the label and the motion field in the previous frame. The main reason is the computational cost that a joint estimation in more than one frames involves. However, such an approach, to our opinion, could disambiguate the estimation of the label and motion field in problematic areas, such as occlusions.

Non rigid motion patterns This includes motion patterns that differ significantly from the usual assumption of local rigidity (e.g. the water of a fountain). Techniques inspired from texture segmentation might be able to provide answers to such situations.

Total occlusions This involves the total disappearance of an object from the scene for a number of frames. Although our methods do not address this issue, there are a number of tracking methods in the literature that do so by preserving the motion hypotheses associated with it.

6.4 Higher level user interaction

Such higher level interaction comes to bridge the semantic gap between the content-based segmentation that a user wishes to obtain and the low level features that can be extracted from an image sequence. Bridging such a gap is crucial in applications such as video editing or video annotation, where the user wants to access and manipulate the contents of the presented material. A system that aims at similar types of applications need to accurately label objects with possibly high degree of variation in their properties in the absence of domain knowledge. Since homogeneity in low level features, such as motion, cannot provide a unique discrimination between the objects, user interaction is necessary. The characteristics of such a system can be summarized as follows:

Absence of domain knowledge The domain and the contents of the video are considered unknown. Therefore, a high degree of freedom in the properties of the expected objects should be allowed.

Satisfactory usable user interface The user interface should be easy to use, intuitive and simple.

Accurate object localization The level of accuracy depends on the specific application but, in general, in such applications that involve user interaction it is expected to be quite high.

The method that we proposed in chapter 5 falls in this category of methods. It proposes the modeling of the **local** statistical properties of the objects and the initialization of the models via a user interaction phase. We do not utilize domain knowledge but employ rather general models that do not make assumptions about the contents of the scene. More specifically, the local modeling that we have proposed implies local smoothness assumptions in color and motion. We have clearly demonstrated that the proposed method is capable of tracking non-rigid as well as rigid motion and objects even when motion information alone is insufficient to discern between them. On the other hand, the smoothness assumption causes the merging of small and elongated areas. This is also due to the fact that the independently estimated motion field in such areas is quite unreliable.

As far as the user interaction is concerned, we proposed a simple and intuitive scheme. The user-specified scribbles initiate a procedure in which borders between objects are declared at the points of highest color gradient between the user's scribbles. Such an approach, attempts to derive an object-based segmentation initiated from the user's scribbles without estimating first the parameters of the models that describe the object properties. The latter is not feasible in our modeling which requires that the label field is known before the model parameters can be estimated. However, even if the modeling would allow it, such an approach has the drawback that it requires the careful initialization of the scribbles such that the object properties along the scribble are sufficiently similar to the object properties. Since we chose to estimate the label field without first estimating the parameters of the models that describe the object properties, we have to utilize homogeneity criteria.

There are a number of alternative ways to utilize homogeneity criteria in order to obtain the initial label field via user interaction. An example of such an alternative is to use a conservative segmentation based on color and/or motion properties. Once a number of segments are extracted the user can specify the collection that comprises the object he is interested in. Such an interaction should include the potential break-up of the initial segments. Evaluation and comparison of our user interface with similar ones in terms of quality metrics (functionality, usability, performance etc) would be an interesting issue for further investigation.

In terms of the accuracy of the label fields the local modeling that we proposed performed well for a number of image sequences. However, the generality of the adopted modeling in the presence of complex scenes reveals a limitation on the proposed method. More specifically, in presence of complex objects whose parts enter and leave the scene the proposed modeling fails to provide a correct labeling. That is an inherent limitation of methods that adopt so general models and, to our opinion, indicate the need for a higher degree of user interaction. In order to facilitate the latter, quality measures should be defined that indicate failures or ambiguities in the tracking. Such measures could result from the objective measure that we optimize or alternative from models on the temporal evolution of the parameters that describe the object properties. Issues that are related to the degree and type of user interaction in such cases need therefore further investigation. Such issues are particularly important for the design of such an interactive system.

Appendix A

Proof of Lemma 1

We will prove that if $(\hat{\Theta}, \hat{d}) = \arg\min\limits_{(\Theta,d)} C_e\left(d, \Theta\right)$ then $\hat{\Theta} = \arg\min\limits_{\Theta} E(L, \Theta, I, \hat{L}^-, I^-, I^+)$.

Clearly if $(\hat{\Theta}, \hat{d}) = \arg\min\limits_{(\Theta,d)} C_e\left(d, \Theta\right)$ then the optimum **direction** field $\hat{d}$ is given by:

$$\hat{d}_s = \begin{cases} 1 & \text{if } \sum_{\mathbf{i}\in G_s}\left(f_{\mathbf{i}}^+\left(\hat{\theta}_{l_s}\right)\right)^2 \leq \sum_{\mathbf{i}\in G_s}\left(f_{\mathbf{i}}^-\left(\hat{\theta}_{l_s}\right)\right)^2 \\ 0 & \text{otherwise} \end{cases} \tag{A.1}$$

Then eq. 4.16 and eq. A.1 imply that:

$$\begin{aligned} C_e\left(\hat{d}, \hat{\Theta}\right) &= \sum_{s=1}^{K} \min\left(\sum_{\mathbf{i}\in G_s}\left(f_{\mathbf{i}}^-\left(\hat{\theta}_{l_s}\right)\right)^2, \sum_{\mathbf{i}\in G_s}\left(f_{\mathbf{i}}^+\left(\hat{\theta}_{l_s}\right)\right)^2\right) \\ &+ E_t(L, \hat{\Theta}, \hat{L}^-) \\ &= E_d\left(I, L, \hat{\Theta}, I^-, I^+\right) + E_t(L, \hat{\Theta}, \hat{L}^-) \end{aligned} \tag{A.2}$$

Let us now consider an arbitrary Θ and denote with $\tilde{d}$ the direction field given by eq. 4.19. Then eq. 4.16 yields:

$$C_e\left(\tilde{d}, \Theta\right) = E_d\left(I, L, \Theta, I^-, I^+\right) + E_t(L, \Theta, \hat{L}^-) \tag{A.3}$$

Then the assumption that $(\hat{\Theta}, \hat{d}) = \arg\min\limits_{(\Theta,d)} C_e\left(d, \Theta\right)$ implies that:

$$\begin{aligned} \forall\Theta, \forall d \qquad & C_e\left(d, \Theta\right) \geq C_e\left(\hat{d}, \hat{\Theta}\right) \Rightarrow \\ \forall\Theta \qquad & C_e\left(\tilde{d}, \Theta\right) \geq C_e\left(\hat{d}, \hat{\Theta}\right) \Rightarrow \\ \forall\Theta \qquad & C_e\left(\tilde{d}, \Theta\right) + E_c(L) \geq C_e\left(\hat{d}, \hat{\Theta}\right) + E_c(L) \Rightarrow \\ \forall\Theta \qquad & E(L, \Theta, I, \hat{L}^-, I^-, I^+) \geq E(L, \hat{\Theta}, I, \hat{L}^-, I^-, I^+) \end{aligned}$$

That is, $\hat{\Theta} = \arg\min\limits_{\Theta} E(L, \Theta, I, \hat{L}^-, I^-, I^+)$

Appendix B

$|G_s| = 1$

In this appendix we will present the degenerate situation of the method proposed in chapter 4 that results when the initial intensity segmentation algorithm provides segments that consist of a single pixel. We will show that in this special case our method reduces to classical MRF formulations for iterative motion estimation / segmentation.

More specifically let us assume that the pixel set of each segment s contains exactly one pixel, that is

$$G_s = \{\mathbf{i}\} \tag{B.1}$$

In that case eq. 4.13 becomes

$$\begin{aligned} E(L,\Theta,I,\hat{L}^-,I^-,I^+) &= E_d(I,L,\Theta,\hat{L}^-,I^-,I^+) + E_c(L) + E_t(L,\Theta,\hat{L}^-) \\ &= \sum_{s=1}^{K} \min\left(\left(f_{\mathbf{i}}^-\left(\theta_{l_s}\right)\right)^2, \left(f_{\mathbf{i}}^+\left(\theta_{l_s}\right)\right)^2\right) \\ &\quad + \sum_{s=1}^{K} \sum_{s' \in N_s} V_c\left(s,s'\right) + \sum_{s=1}^{K} V_{ts}\left(\hat{L}^-, s, \theta_{l_s}\right) \end{aligned} \tag{B.2}$$

Since there is no longer a distinction between segments and pixels, in eq. B.2 the indexes s and $\mathbf{i}$ bear the same meaning. For consistency in the terminology we preserve the “segment”-like notation.

The first term of eq. B.2 is the classical observation term which corresponds to the modeling of the noise along the motion trajectories. Our model, where backward and forward motion-compensated intensity differences are considered, is similar to that of Dubois and Konrad[32], where visibility sets are defined.

The second term in eq. B.2 expresses the spatial interactions between the pixels. The region adjacency graph has been reduced to the regular image lattice and the cliques are defined as pairs of neighboring pixels. The length of the common border $b(s,s')$ of two neighbors is obviously one, so the clique potential (eq. 4.12) becomes:

$$V_c\left(s,s'\right) = \begin{cases} -z_c & \text{if } l_s = l_{s'} \\ z_c & \text{if } l_s \neq l_{s'} \end{cases} \tag{B.3}$$

This is a classical Markovian potential which favors smooth label fields and penalizes large frontiers. Note that the smoothing is independent of the local image structure. That is because the formulation of eq. 4.12 assumes that pixels with low intensity gradient are already grouped in a single segment. However, one can think of meaningful definitions of clique potentials at segment level that introduce anisotropic spatial constraints in the degenerate case.

Finally the third term in eq. B.2 is term which favors temporal continuity of the label field along motion trajectories. Stiller [104] defines a similar constraint, but he favors temporally consistent labeling of cliques and not of each pixel separately.

As far as the optimization procedure in the degenerate case is concerned, our method would fall naturally in the area where iterative algorithms are used in the MAP-MRF framework. For example the algorithm that Chang *et al.* [23] propose also involves a scheme where external and internal iterations are considered.

Appendix C

Relation with the EM algorithm

In this appendix we will prove that the method proposed in chapter 4 belongs to the class of the *Expectation Maximization* algorithms that employ hard decisions. Let us formulate the problem in the EM framework. Adopting the EM terminology, the motion parameters Θ are the parameters to be estimated, the image intensities I, I^+, I^- and the estimation of the label field L^- are the *observed* data and the label field L is the *latent* data. Then, the conditional probability of the *complete* data is:

$$P\left(I, I^+, I^-, \hat{L}^-, L|\Theta\right) = P\left(I|L, \Theta, \hat{L}^-, I^-, I^+\right) \\ P\left(I^+, I^-|L, \Theta, \hat{L}^-\right) P\left(\hat{L}^-|L, \Theta\right) P\left(L|\Theta\right) \quad \text{(C.1)}$$

Note that eq. C.1 is identical to eq. 4.3, except of the second term on the right-hand side. That term expresses the dependencies between the intensities of the previous and the next frame, which are ignored in our formulation. However, such dependencies are in general more sensitive to occlusions and it is questionable if their incorporation would have a significant positive contribution.

In the *Expectation* step of the EM algorithm the goal is to find the Expectation of the negative log likelihood of the *complete data*. Under our assumptions for the conditional probabilities, that is equal to:

$$Q(\Theta|\Theta_m) = \sum_{s=1}^{K} \sum_{n=1}^{N} g_{sn} \left(V_{ds} + V_{ts}\right) + \mathcal{E}\left\{E_c\right\} - Z_1 - Z_2 - Z_3 \quad \text{(C.2)}$$

where $g_{sn} = P\left(l_s = n|I, I^+, I^-, \hat{L}^-, \Theta\right)$ and Z_1, Z_2 and Z_3 are normalization constants in the Gibbs distributions. In order to proceed to the Maximization step we only need to estimate g_{sn}, since only the terms V_{ds} and V_{ts} are dependent on Θ. However, this is not trivial since the MRF modeling of the label field generates dependencies on the conditional probabilities of the different segments. In order to overcome this, a common strategy [11][120] is to approximate $P(l_s = n)$ by considering as known estimates of the labels of the neighboring segments. These estimates denoted by $\hat{l}_{s'} : s' \in N_s$ are provided at the intermediate steps of an iterative scheme. That is:

$$P(l_s = n) \approx P(l_s = n | \hat{l}_{s'} : s' \in N_s) \tag{C.3}$$

Then we can easily derive g_{sn} as:

$$g_{sn} = \frac{e^{-\left(V_{ds} + V_{ts} + \sum_{s' \in N_s} V_c(s,s') | l_s = n, \hat{l}_{s'}\right)}}{\sum_{n'=1}^{N} e^{-\left(V_{ds} + V_{ts} + \sum_{s' \in N_s} V_c(s,s') | l_s = n', \hat{l}_{s'}\right)}} \tag{C.4}$$

and by this the *Expectation* step is complete.

In the EM framework with hard decisions, the label n with the higher conditional probability g_{sn} is chosen. This is exactly the same as the step 3 in the iterative Labeling phase that we employ (Table 4.1).

Finally, it is straightforward to show that the Maximization step where the maximization of $Q(\Theta|\Theta_m)$ with respect to Θ takes place, is equivalent to our Motion Estimation phase. Let us note here that with the above formulation the different g_{sn} could be used in the motion estimation phase. This would result in an EM algorithm with soft decisions.

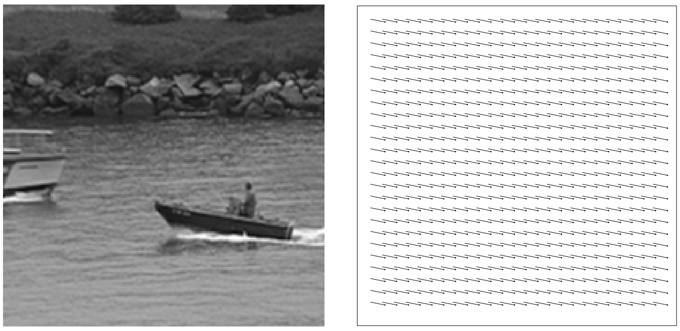
Appendix D

Synthetic Image Sequences

This appendix contains a short description of the synthetic image sequences that are used in chapter 3. Each sequence consists of three frames. Since the model-generated motion fields are real-valued, a bicubical interpolator on the image intensities was used.

D.1 “C1” image sequence

Translational Motion of the Background An image sequence is generated, in which the whole image is displaced by $(5, 1)$ pixels per frame. The second frame of the sequence is depicted in fig. D.1(a) and the model-generated motion field (magnified by a factor of 2) in fig. D.1(b).



(a) Frame 2

(b) Model-generated motion field

Figure D.1: “C1” image sequence

D.2 “R1” image sequence

Translational Motion of the Background An image sequence is generated, in which the whole image is displaced by $(1, 1)$ pixels per frame. The second frame of the sequence is depicted in fig. D.2. The main characteristic of the sequence is the lack of texture in large areas.

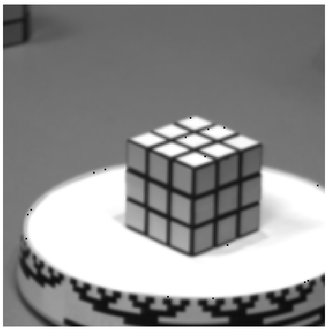


Figure D.2: Frame 2 of “R1” image sequence

D.3 “Y1” image sequence

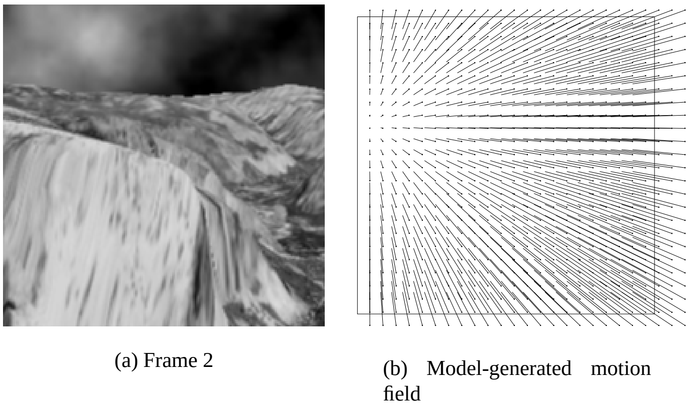
Affine Motion of the Background An image sequence is generated in which the background is displaced according to an affine parametric model (Table D.1). The second frame is depicted in fig. D.3(a) and the model-generated motion field (magnified by a factor of 2) in fig. D.3(b).

	$\theta(1)$	$\theta(2)$	$\theta(3)$	$\theta(4)$	$\theta(5)$	$\theta(6)$
Background	0.1	0	10	0	0.1	3

Table D.1: Motion parameters for “Y1” image sequence

D.4 “S5” image sequence

Two objects, large affine motions An image sequence is generated in which the background and an object are displaced according to two different affine parametric models (Table D.2). The affine parameters are chosen such that the magnitude of motion is quite large, thus large occlusions are present. Due to occlusion phenomena, areas on the left and on the right of the boat are visible only in the second frame of the sequence. The second frame is depicted in fig. D.4(a) and the model-generated motion field in fig. D.4(c). The object’s mask in the second frame is depicted in fig. D.4(b).

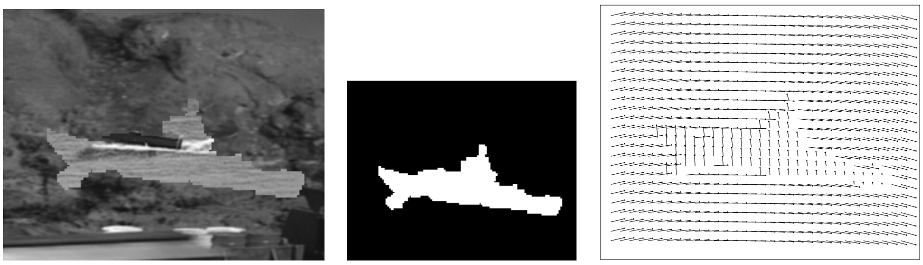

(a) Frame 2

(b) Model-generated motion field

Figure D.3: "Y1" image sequence

	$\theta(1)$	$\theta(2)$	$\theta(3)$	$\theta(4)$	$\theta(5)$	$\theta(6)$
Background	0.01	-0.005	25	0.05	0	1
Object	0	0.03	-2	0.1	0	-13

Table D.2: Motion parameters for "S5" image sequence


(a) Frame 2

(b) Object mask

(c) Model-generated motion field

Figure D.4: "S5" image sequence

Appendix E

Terminology

$\mathcal{E}\left\{y\right\}$	The expected value of a random variable y
α	Parameter controlling the degree of locality of the color-motion model
β	The coefficient that is assumed to relate $\lambda_{B_{\mathbf{i}}}$ and $\sigma_{B_{\mathbf{i}}}$ (i.e. $\lambda_{B_{\mathbf{i}}} = \frac{\beta}{\sigma_{B_{\mathbf{i}}}}$)
$\Theta = \{\theta_1, \ldots, \theta_N\}$	The set of the parameters of the motion models for all of the N objects
$\theta_n : n \in [1 \ldots N]$	Parameters of the motion model for object n
θ_{sn}	The parameters of the Gaussian distribution around segment s for the object with label n
$\lambda_{B_{\mathbf{i}}}$	Deviation of the Laplacian which models the distribution of the motion-compensated intensity differences in block $B_{\mathbf{i}}$
$\mu_d(\phi)$	The (scalar) mean of the d-th component of the multivariate Gaussian with parameters ϕ
π_{sn}	The local *a priori* probability of the object n
$\sigma_{B_{\mathbf{i}}}$	Standard deviation of the intensity within block $B_{\mathbf{i}}$
σ_c	Scale parameter of the Geman McClure that was used for motion clustering
ϕ_s	The parameters (mean and covariance) of the multivariate Gaussian that models the distribution of the feature vectors for the pixels in color segment s
$B^h_{\mathbf{i}}$	The set of pixels that belong to the ancestor at level h of the block centered at pixel $\mathbf{i}$
C_s	Rectangular support region defined around color segment s
$DC(\mathbf{i}, \mathbf{j})$	Color difference between pixel $\mathbf{i}$ in current frame and $\mathbf{j}$ in the previous frame
$D^h_{\mathbf{i}}(\mathbf{v}^h)$	Mean-absolute displaced block difference for block $B^h_{\mathbf{i}}$

$f_{\mathbf{i}}^{-}(\theta)$	Backward motion-compensated intensity difference under the motion hypothesis θ at pixel $\mathbf{i}$
$f_{\mathbf{i}}^{+}(\theta)$	Forward motion-compensated intensity difference under the motion hypothesis θ at pixel $\mathbf{i}$
G_s	Set of pixels of intensity(color) segment s
H	Height of the multiscale pyramid
h	A level in the multiscale scheme
$I(\mathbf{i})$, $I^{-}(\mathbf{i})$ and $I^{+}(\mathbf{i})$	Image intensity at pixel $\mathbf{i}$ at current, previous and next frame, respectively
$\mathbf{i} = (\mathbf{i}_x \; \mathbf{i}_y)$	A pixel as the pair of its coordinates
$L = \{l_s : s \in [1 \ldots K]\}$	Label field
$l_{\mathbf{i}}$	Label of pixel $\mathbf{i}$. Where appropriate, $l_{\mathbf{i}} = l_s : \mathbf{i} \in G_s$
$l_s \in [1 \ldots N]$	Label of intensity(color) segment s
K	The number of intensity(color) segments
N	The number of objects
N_s	The set of the neighbors of segment s
n	The index of an object
S_s	The set of segments that intersect with the support region C_s of color segment s
$s \in [1 \ldots K]$	Intensity(color) segment index
$\hat{\mathbf{v}}_{\mathbf{i}}$	Estimated motion vector at the block $B_{\mathbf{i}}$
$\tilde{\mathbf{v}}_{\mathbf{i}}$	Model-generated motion vector at pixel $\mathbf{i}$
$X = \{\mathbf{x}_{\mathbf{i}}\}$	The set of the feature vectors
$\overline{y}$	The estimated mean value of a random variable y

Bibliography

[1] G. Adiv. Determining 3-d motion and structure from optical flow generated by several moving objects. *PAMI*, 7(4):384–401, July 1985.

[2] G. Adiv. Inherent ambiguities in recovering 3-d motion and structure from a noisy flow field. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 11(5):477–489, May 1989.

[3] Y. Altunbasak, Eren P.E., and A.M. Tekalp. Region-based parametric motion segmentation using color information. *Graphical models and Image Processing*, 60(1):13–23, Jan 1998.

[4] P. Anandan, J. Bergen, K. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In M. Sezan and R. Lagendijk, editors, *Motion Analysis and Image Sequence Processing*. Kluwer Academic Press, 1993.

[5] A.P.Dempster, N.M.Laird, and D.B.Rubin. Maximum likelihood from incomplete data via the *em* algorithm. *Journal Royal Statistical Society, Series B.*, 1:1–38, 1977.

[6] Serge Ayer. *Sequential and competitive methods for estimation of multiple motions*. PhD thesis, Ecole Polytechnique Federale de Lausanne, 1995.

[7] J.L. Barron, D.J. Fleet, and S.S. Beauchemin. Performance of optical flow techniques. *Int'l Journal Computer Vision*, 12(1):43–77, February 1994.

[8] A. Baumberg and D. Hogg. Learning deformable models for tracking the human body. In M. Shah and R. Jain, editors, *Motion-Based Recognition*, Computational Imaging and Vision, pages 39–60. Kluwer Academic Publishers, 1997.

[9] M. Bertero, T. A. Poggio, and V. Torre. Ill-posed problems in early vision. *Proc. of IEEE*, 76(8):869–889, Aug. 1988.

[10] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society*, B-36:192–236, 1974.

[11] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society*, 48(3):259–302, 1986.

[12] S. Beucher. Watersheds of functions and picture segmentation. In *Proc. IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pages 1928–1931, May 1982. Paris, France.

[13] M. Bierling. Displacement estimation by hierarchical block matching. In T.R. Hsing, editor, *Proc. SPIE Visual Communications and Image Processing*, volume 1001, pages 942–951, 1988.

[14] M.J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow-fields. *Computer Vision and Image Understanding*, 63(1):75–104, Jan. 1996.

[15] M.J. Black and A. Rangarajan. On the unification of line processes, outlier detection and robust statistics with applications in early vision. *Int'l Journal of Computer Vision*, 19(1):57–91, Jan. 1996.

[16] A. Blake, R. Curwen, and A. Zisserman. A framework for spatio-temporal control in the tracking of visual contour. *Int'l Journal of Computer Vision*, 11(2):127–145, 1993.

[17] G.D. Borshukov, G. Bozdagi, Y. Altunbasak, and A.M. Tekalp. Motion segmentation bu multi-stage affine classification. *IEEE Trans. Image Processing*, 6(11):1591–1594, Nov. 1997.

[18] P. Bouthemy and E. Francois. Motion segmentation and qualitative dynamic scene analysis from an image sequence. *Int'l Journal of Computer Vision*, 10(2):157–182, May 1993.

[19] N. Brady and N.E. O'Connor. Object detection and tracking using an em-based motion estimation and segmentation framework. In *Proc. IEEE Int'l Conf. Image Processing*, page 17A2, 1996.

[20] R.G. Brown and P.Y.C. Hwang. *Introduction to Random Signals and Applied Kalman Filtering*. John Wiley & Sons, 1996. 3rd edition.

[21] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *Int'l Journal of Computer Vision*, 22(9):61–79, 1997.

[22] Edmond Chalom. *Statistical Image Sequence Segmentation Using Multidimensional Attributes*. PhD thesis, Massachusetts Institute of Technology, 1998.

[23] M. M. Chang, A. M. Tekalp, and M. I. Sezan. An algorithm for simultaneous motion estimation and scene segmentation. In *Proc. IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, Apr. 1994. Adelaide, Australia.

[24] G.C Choi and S. Kim. Multi-stage segmentation of optical flow field. *Signal Processing*, 54:109–118, 1996.

[25] P.B. Chou and C.M. Brown. The theory and practice of bayesian image labeling. *Int'l Journal of Computer Vision*, 4(3):185–210, 1990.

[26] Sarnoff Corporation. Multimedia Composition and Authoring Toolkit. http://www.sarnoff.com.

[27] T. Darrell and A.P. Pentland. Cooperative robust estimation using layers of support. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 17(5):474–487, May 1995.

[28] R. Dave and R. Krishnapuram. Robust clustering methodes: A unified view. *IEEE Trans. Fuzzy Systems*, 5(2):270–293, Feb. 1997.

[29] H. Derin and H. Elliot. Modeling and segmentation of noisy and textures images using gibbs random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9:39–55, Jan. 1987.

[30] A. Dev, B.J.A. Krose, and F.C.A. Groen. Confidence measures for image motion estimation. In *RWC Symposium*, pages 199–206, 1997.

[31] N. Diehl. Object-oriented motion estimation and segmentation in image sequences. *Signal Processing: Image Communications*, 3(1):23–56, Jan. 1991.

[32] E. Dubois and J. Konrad. *Estimation of 2-D Motion Fields from Image Sequences with Application to Video Coding*, pages 53–87. In Motion Analysis and Image Sequence Processing. M.I. Sezan and R.L. Lagendijk, eds., Kluwer Academic Publishers, 1993.

[33] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, 1973.

[34] F. Dufaux, F. Moscheni, and A. Lippman. Spatiotemporal segmentation based on motion and static segmentation. In *Proc. IEEE Int'l Conf. Image Processing*, volume 1, pages 306–309, Oct. 1995. Washington, DC.

[35] R. Fablet, P. Bouthemy, and M. Gelgon. Moving object detection in color image sequences using region-level graph labeling. In *Proc. IEEE Int'l Conf. Image Processing*, Oct. 1999. Kobe, Japan.

[36] H. Frigui and R. Krishnapuram. A robust algorithm for automatic extraction of an unknown number of clusters from noisy data. *Pattern Recognition Letters*, 17:1223–1232, 1996.

[37] C. S. Fuh and P. Maragos. Affine models for image matching and motion detection. Technical Report CICS-P-280, Center for Intelligent Control Systems, Feb. 1991.

[38] N. P. Galatsanos and A. K. Katsaggelos. Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation. *IEEE Trans. Image Processing*, 1(3):322–336, Jul. 1992.

[39] D. M. Gavrilla. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, Jan. 1999.

[40] M. Gelgon and P. Bouthemy. A region-level graph labeling approach to motion-based segmentation. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 514–519, 1997.

[41] M. Gelgon and P. Bouthemy. A region-level motion-based graph representation and labeling for tracking a spatial image partition. *Pattern Recognition*, 33:725–740, Apr. 2000.

[42] S. Geman and D. Geman. Stochastic relaxation, gibbs distribution and the bayesian restoration of images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 6:721–741, Nov. 1984.

[43] S. Ghosal and P. Vanek. A fast scalable algorithm for discontinuous optical-flow estimation. *PAMI*, 18(2):181–194, February 1996.

[44] P. Golland and A. M. Bruckstein. Motion from color. *Computer Vision and Image Understanding: CVIU*, 68(3):346–362, December 1997.

[45] Chuang Gu. *Multivalued Morphology and Segmentation-based Coding*. PhD thesis, Ecole Polytechnique Federale de Lausanne, 1995.

[46] P.W. Holland and R.F. Welsch. Robust regression using iteratively reweighted least squares. *Comm. Statistics. Theory and Methods*, 1977.

[47] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1-3):185–203, Aug. 1981.

[48] Y. Huang, K. Palaniappan, X. Zhuang, and J.E. Cavanaugh. Optic flow field segmentation and motion estimation using a robust genetic partitioning algorithm. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 17(12):1177–1190, Dec. 1995.

[49] M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *Int'l Journal of Computer Vision*, 29(1):5–28, 1998.

[50] I. Kakadiaris and D. Metaxas. Model-based estimation of 3-d human motion. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(12):1453–1459, Dec. 2000.

[51] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An introduction to Cluster Analysis*. Wiley, New York, 1990.

[52] R. Keys. Cubic convolution interpolation for digital image processing. *IEEE Trans. Acoustics, Speech and Signal Process.*, 29:1153–1160, Dec. 1981.

[53] S. Kirpatrick, C.D. Gelatt, Jr., and M.P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, May 1983.

[54] J. Konrad and V.N. Dang. Coding-oriented video segmentation inspired by mrf models. In *Int'l Conf. Image Processing*, volume 1, pages 909–912, 1996.

[55] J. Konrad and E. Dubois. Bayesian estimation of motion vector fields. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 14(9):910–927, Sep. 1992.

[56] J. Konrad and C. Stiller. *On Gibbs-Markov Models for Motion Computation*, pages 121–154. In Video Data Compression for Multimedia Computing: Statistically Based and Biologically Inspired Techniques. H. Li, S. Sun and H. Derin, eds., Kluwer Academic Publishers, 1997.

[57] S.Z. Li. Markov random field modeling in computer vision. *Springer-Verlag*, pages ISBN 0–387–70145–1, 1995.

[58] S.Z. Li. Toward global solution to map image restoration and segmentation: Using common structure of local minima. *Pattern Recognition*, 33(4):715–723, Apr. 2000. URL: http://markov.eee.ntu.edu.sg:8000/ szli/publications.html.

[59] D. G. Lowe. Robust model-based motion tracking through the integration of search and estimation. *Int'l Journal of Computer Vision*, 2(8):113–122, 1993.

[60] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Int'l Joint Conference on Artificial Intelligence*, pages 121–130, 1981.

[61] A. Lundmark, H. Li, and R. Forchheimer. Motion vector certainty reduces bit rate in backward motion estimation video coding. In *SPIE Visual Communications and Image Processing*, Jun. 2000.

[62] A. Mansouri and Konrad J. Multiple motion segmentation with level sets. *IEEE Transactions on Image processing*, 2000. SUBMITTED.

[63] P. Meer, D. Mintz, and A. Rosenfeld. Robust regression methods for computer vision: A review. *Int'l Journal of Computer Vision*, 6(1):59–70, 1991.

[64] F.G. Meyer. Color image segmentation. In *Proc. IEEE Int'l Conf. Image Processing and its applications*, pages 303–304, May 1992. Maastricht, The Netherlands.

[65] F.G. Meyer and S. Beucher. Morphological segmentation. *Journal of Visual Communication and Image Processing*, 1(1):21–46, Sep. 1990.

[66] F.G. Meyer and P. Bouthemy. Region-based tracking using affine motion models in long image sequences. *CVGIP: Image Understanding*, 60(2):119–140, Sep. 1994.

[67] F. Moscheni, S. Bhattacharjee, and M. Kunt. Spatiotemporal segmentation based on region merging. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(9):897–915, Sep. 1998.

[68] F. Moscheni, F. DuFaux, and M. Kunt. A new two-stage global/local motion estimation based on a background/foreground segmentation. In *Int'l Conf. Acoustics, Speech and Signal Processing*, May 1995. Detroit, MI.

[69] Moving Picture Experts Group. http://www.cselt.it/mpeg.

[70] MPEG. MPEG-4: Applications document. Technical Report ISO/IEC JTC1/SC29/WG11 w2724, MPEG, Mar. 1999. Seoul, Korea. http://www.cselt.it/mpeg/public/mpeg-4_applications.zip.

[71] MPEG. MPEG-4: Requirements document. Technical Report ISO/IEC JTC1/SC29/WG11 N3930, MPEG, Jan. 2001. Pisa. http://www.cselt.it/mpeg/public/mpeg-4_requirements.zip.

[72] A. Vetro Ed. MPEG. MPEG-7: Applications document. Technical Report ISO/IEC JTC1/SC29/WG11/N3934, MPEG, Jan. 2001. Pisa. http://www.cselt.it/mpeg/public/mpeg-7_applications.zip.

[73] F. Pereira Ed. MPEG. MPEG-7: Requirements document. Technical Report ISO/IEC JTC1/SC29/WG11/N4035, MPEG, Mar. 2001. Singapore. http://www.cselt.it/mpeg/public/mpeg-7_requirements.zip.

[74] David W. Murray and Bernard F. Buxton. Scene segmentation from visual motion using global optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9(2):220–228, Mar 1987.

[75] H.G. Musmann, P. Pirsch, and H.J. Grallert. Advances in image coding. *Proc. of IEEE*, 73:523–548, Apr. 1985.

[76] H.H. Nagel. On the estimation of optical flow: Relations between different approaches and some new results. *Artificial Intelligence*, 33(3):299–324, Nov. 1987.

[77] H.H. Nagel and W. Enkelmann. An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 8(5):565–593, Sep. 1986.

[78] R.M. Neal and G.E. Hinton. A new view of the EM algorithm that justifies incremental, sparse, and other variants. In M.I. Jordan, editor, *Learning in Graphical Model*. 1998.

[79] H.T. Nguyen and M. Worring. Multifeature object tracking using a model-free approach. In *IEEE conference on Computer Vision and Pattern Recognition*, volume 1, pages 145–150, 2000. Hilton Head, USA.

[80] S. Nitsuwat and J. S. Jin. Analysing motion parameters using unsupervised fuzzy c-prototypes. In *Pan-Sydney Area Workshop on Visual Information Processing*, pages 27–33, 1998. Sydney, Australia.

[81] S. Nitsuwat, J. S. Jin, and H.M. Hudson. Motion-based video segmentation using fuzzy clustering and classical mixture model. In *International Conference on Image Processing*, volume 1, pages 300–303, 2000. Vancouver, Canada.

[82] Noel E. O'Connor, Noel Brady, and Sean Marlow. Supervised image segmentation using em-based estimation of mixture density parameters. In *Workshop on Image Analysis for Multimedia Interactive Services*, pages 27–32, Jun. 1997.

[83] J. M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Journal of Visual Communication and Image Representation*, 6(4):348–365, Dec. 1995.

[84] Parmenides of Elea. *On Nature (Peri Physeos)*. c. 475 B.C.

[85] Nuria Oliver, Alex Pentland, and Francois Berard. Lafter: Lips and face real time tracker with facial expression recognition. In *IEEE conference on Computer Vision and Pattern Recognition*, pages 123–129, Jun. 1997.

[86] M. Pantic and L.J.M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, Dec. 2000.

[87] Thrasyvoulos N. Papas. An adaptive clustering algorithm for image segmentation. *IEEE Trans. Signal Processing*, 40(4):901–914, Apr. 1992.

[88] A. Papoulis. *Probability, Random Variable and Stochastic Processes*. Electrical and Electronic Engineering series. McGraw-Hill, 1991. 3rd edition.

[89] N. Paragios and R. Deriche. A pde-based level set approach for detection and tracking of moving objects. In *International Conference on Computer Vision*, pages 1139–1145, Jan. 1998. Bombay, India.

[90] N. Paragios and G. Tziritas. Adaptive detection and localization of moving objects in image sequences. *Signal Processing: Image Communication*, 14(4):278–296, Feb. 1999.

[91] I. Patras, E.A. Hendriks, and R.L. Lagendijk. Segmentation of image sequences applying mrf on watershed segments. In *IEEE Benelux Signal Processing Symposium*, pages 147–150, Mar. 1998. Leuven, Belgium.

[92] I. Patras, E.A. Hendriks, and R.L. Lagendijk. A semi-automatic method for segmentation of image sequences with local region-based classification. In *IASTED International Conference on Signal and Image Processing*, Nov. 2000. Las Vegas, USA.

[93] I. Patras, E.A. Hendriks, and R.L. Lagendijk. Video segmentation by map labeling of watershed segments. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(3):326–332, Mar. 2001.

[94] N. Peterfreund. Robust tracking of position and velocity with kalman snakes. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 21(6):564–569, Jun. 1999.

[95] J. M. Rehg and T. Kanade. Visual tracking of high DOF articulated structures: An application to human hand tracking. *Lecture Notes in Computer Science*, 800:35–46, 1994.

[96] J Rissanen. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 11(2):416–431, 1983.

[97] S. Harpreet Sawhney and Serge Ayer. Compact representation of videos through dominant and multiple motion estimation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18(8):814–830, Aug 1996.

[98] R.J. Schalkoff and E.S. McVey. A model and tracking algorithm for a class of video targets. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 4(1):2–10, Jan. 1982.

[99] J.A. Sethian. *Level Set Methods*. Cambridge University Press, 1996.

[100] L. Shafarekno, M. Petrou, and J. Kittler. Automatic watershed segmentation of randomly textured color images. *IEEE Transactions on Image processing*, 6(11):1530–1544, Nov. 1997.

[101] P. Shalembier and J. Serra. Morphological multiscale image segmentation. In *Proc. SPIE Visual Communications and Image Processing*, volume 1818, pages 620–631, Nov. 1992. Boston, Massachusetts.

[102] E.P. Simoncelli. *Distributed Representation and Analysis of Visual Motion*. PhD thesis, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, 1993.

[103] E.P. Simoncelli, E.H. Adelson, and D.J. Heeger. Probability distributions of optical flow. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 310–315, June 1991. Maui, Hawaii.

[104] C. Stiller. Object-based estimation of dense motion fields. *IEEE Trans. Image Processing*, 6(2):234–250, Feb. 1997.

[105] C. Stiller and J. Konrad. Estimating motion in image sequences: A tutorial on modeling and computation of 2d motion. *IEEE Signal Processing Magazine*, 16:70–91, Jul. 1999.

[106] R. Szeliski and D. Terzopoulos. Physically-based and probabilistic modeling for computer vision. In *Proc. SPIE 1570, Geometric Methods in Computer Vision*, pages 140–152, Jul. 1991. San Diego, USA.

[107] H. Tao, H.S. Sawhney, and R. Kumar. dynamic layer representation and its applications to tracking. In *IEEE conference on Computer Vision and Pattern Recognition*, Jun. 2000. Hilton Head.

[108] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, 1999.

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

[110] L. Vincent and P. Soille. Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13:583–589, Jun. 1991.

[111] C.S. Wallace and D.M. Boulton. An information measure for classification. *Computing Journal*, 11(2):185–195, 1968.

[112] D. Wang. Unsupervised video segmentation based on watersheds and temporal tracking. *IEEE Trans.s in Circuits and Systems for Video Technology*, 8(5):539–546, Sep. 1995.

[113] J. Y. A. Wang and E. H. Adelson. Representing moving images with layers. *IEEE Trans. Image Processing*, 3(5):625–638, Sep. 1994.

[114] J.Y.A. Wang and E.H. Adelson. Layered representation for motion analysis. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 361–366, 1993.

[115] Y. Weiss and E.H. Adelson. A unified mixture framework for motion segmentation: Incorporating spatial coherence and estimating the number of models. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 321–326, 1996.

[116] Yair Weiss and Edward H. Adelson. Perceptually organized em: A framework for motion segmentation that combines information about form and motion. In *International Conference on Computer Vision*, 1995.

[117] T. Yoshida, H. Katoh, and Y. Sakai. Block matching motion estimation using block integration based on reliability metric. In *Int'l Conf. Image Processing*, volume 2, pages 152–155, 1997.

[118] T. Y. Young, W. Z. Zhao, Qi F. H., and D. Ergener. Computation of image velocity field using linear and nonlinear objective function. In *Proc. IEEE Workshop on Computer Vision*, pages 342–344, Dec. 1987.

[119] F. Zanoguera, B. Marcotegui, and F. Meyer. A toolbox for interactive segmentation based on nested partitions. In *Proc. IEEE Int'l Conf. Image Processing*, 1999.

[120] J. Zhang, J.W. Modestino, and D.A. Langan. Maximum-likelihood parameter estimation for unsupervised stochastic model-based image segmentation. *IEEE Trans. Image Processing*, 3(4):404–419, Jul. 1994.

Samenvatting

In dit proefschrift worden drie methoden voor objectgebaseerde segmentatie van beeldsequenties voorgesteld. In alle drie methoden wordt voor ieder beeld uit de tijdreeks een intensiteit- of kleurensegmentatie uitgevoerd, die een verzameling van fijne segmenten oplevert. Deze segmenten worden vervolgens gelabeld op basis van hun bewegings-, en kleureigenschappen. We tonen de voordelen aan van zo een aanpak in termen van robuustheid, nauwkeurigheid van lokalisatie en computationele complexiteit.

In de eerste twee methoden wordt er van uit gegaan dat een bekend aantal objecten in de scne gedentificeerd kan worden op basis van hun bewegingspatroon. Bovendien wordt aangenomen dat dit patroon beschreven kan worden door geparametriseerde modellen van een lage orde.

De eerste methode (hoofdstuk 3) stelt een benadering voor waarbij het schatten van beweging en het labellen in achtereenvolgende stappen worden uitgevoerd. In de eerste stap wordt een bewegingsveld geschat met behulp van een hirarchische blockmatching algoritme. Dit bewegingsveld wordt vervolgens geclusterd onder de aanname dat het bewegingsveld van elk cluster beschreven kan worden door een affine model. De affine modellen dienen als bewegingshypotheses. Om te kunnen omgaan met onnauwkeurigheden in het bewegingsveld, is een clusteringmethode ontwikkeld die gebruik maakt van bewegings-specifieke betrouwbaarheidsmaten en technieken die genspireerd zijn door methoden uit de robuuste statistiek. In de tweede stap wordt aan elk segment, dat het resultaat is van de initile intensiteitssegmentatie, een objectlabel toegewezen op basis van twee criteria: a) de bewegingshypotheses die het best overeenkomen met het bewegingspatroon van dat segment b) de waarden van het labelveld in de omgeving van dat segment. Deze criteria zijn geformuleerd door a) het modelleren van de distributie van de beweging-gecompenseerde intensiteitverschillen door een Gaussische distributie en b) door het modelleren van het labelveld als een Markov Random Field. De verbindingen ('cliques') in het Markov Random Field zijn gedefinieerd tussen intensiteitsegmenten. Een drie-frame benadering wordt toegepast om met occlusie om te kunnen gaan.

De tweede methode (hoofdstuk 4) drukt de ruimtelijke en temporele voorwaarden van het probleem van labelen uit binnen een zelfde kader en maakt bovendien een gelijktijdige schatting van het labelveld en de parameters van de bewegingsmodellen. Dit gebeurt door het maximaliseren van de *a posteriori* kans van het labelveld. Ruimtelijke en temporele voorwaarden aan het labelveld worden uitgedrukt in het Markov Random Field kader waarin verbindinngen zijn gedefinieerd tussen intensiteitsegmenten. Voor de optimalisatie stellen wij een methode voor die de *a posteriori* kans vergroot op een iteratieve manier met betrekking tot de bewegingsparameters en het labelveld. Ook hier wordt een drie-frame benadering gebruikt om met occlusie om te kunnen gaan. We tonen aan dat een aantal op pixels gebaseerde methoden uit gedrukt kunnen worden als een bijzondere geval van onze methode. Tevens tonen wij aan hoe onze methode uitgebreid kan worden om aan elk intensiteitsegment een objectlabel toe te wijzen met een zekerheid dat gelijk staat aan de overeenkomstige *a-posteriori* kans (soft labeling decisions).

De derde methode (hoofdstuk 5) stelt een semi-automatische benadering voor om meer complexe objecten aan te kunnen pakken die niet altijd of volledig te karakteriseren zijn door hun bewegingspatroon. Voor het eerste frame van de beeldsequentie wordt een beschrijving gemaakt van de lokale statistische eigenschappen van het object. Dit gebeurt op basis van een labelveld dat is genitieerd door krabbels die door de gebruiker zijn gespecificeerd. Vervolgens wordt het labelveld getraceerd in de rest van de beeldsequentie. Het labelen wordt uitgevoerd door een op kansen gebaseerde classificatie van de segmenten die het resultaat zijn van de initile kleurensegmentatie stap. Aangenomen wordt dat de data van punten binnen een bepaald kleurensegmentatie gegenereerd wordt door hetzelfde proces dat gemodelleerd is als een multivariate Gaussische distributie. De voorwaardelijke kans van beweging en kleur, gegeven het labelveld, wordt gemodelleerd binnen een bepaald omgeving rond het middelpunt van ieder segment als een mix van multivariate Gaussische distributies, waarbij elk Gaussische distributie correspondeert met een bepaald object. Het classificatie criterium is de maximalisatie van de gecombineerde kans van het labelveld en de observaties met betrekking tot het labelveld. Voor de maximalisatie van de gecombineerde kans is een deterministische iteratieve lokale zoek algoritme ontwikkeld.

Curriculum Vitae

Ioannis Patras

Ioannis Patras was born in Thessaloniki, Greece, in 1973. In 1990 he obtained his Lyceum diploma from the Third Lyceum in Veroia, Greece. In the same year he entered the Computer Science Department at University of Crete in Heraklion, Greece, from which he received his B.Sc. in 1994. In 1997 he received his M.Sc. degree from the same department with specialization in the areas of i) Machine Vision and Robotics and ii) Parallel and Distributed Systems. From 1996 until 2000 he worked towards his Doctorate degree at Delft University of Technology.

His research experience begins in 1992, when as an undergraduate student he worked in the Machine Learning group of Computer Science Institute in Heraklion, Greece. From 1994 until 1996 he worked in the Computer Vision and Robotics group of the same institute in the field of Dynamic Stereoscopic Vision. Between April 1996 and June 1996 he was a visiting researcher at the Information Theory Group at Delft University of Technology. In December 1996 he joined the same group as Ph.D. researcher in the field of Object-based Segmentation of Image Sequences. In April 2001 he joined the Intelligent Sensory Information Systems group at Computer Science Department at University of Amsterdam in the field of Information Retrieval. He has around 10 publications in international conferences and journals. He has supervised a number of M.Sc. students the work of whom have been acknowledged by publications in national and international conferences.