WP5 : State of the Art Report: Image and Video Processing for Multimedia Understanding

Edited by Nozha Boujemaa[1], Hichem Houissa[1], Horst Bischof[2]
[1]INRIA - Imedia, France
[2]CGV - Graz University of Technology, Austria

www-rocq.inria.fr/imedia/Muscle/WP5

Contents

List of authors

FT R&D, France

Nathalie Laurent	nathalie.laurent@francetelecom.com
Christophe Laurent	christophe2.laurent@francetelecom.com
Mariette Maurizot	mariette.maurizot@wanadoo.fr

ENST, France

Béatrice Pescquet-Popescu	beatrice.pesquet@enst.fr
Gemma Pielle	piella@tsi.enst.fr

INRIA-Vista, France

Patrick Bouthemy	Patrick.Bouthemy@irisa.fr
Frederic Cao	Frederic.Cao@irisa.fr
Gwenaelle Piriou	Gwenaelle.Piriou@irisa.fr
Thomas Veit	Thomas.Veit@irisa.fr
Patrick Perez	Patrick.Perez@irisa.fr

ISTI-CNR Group, Italy

Emanuele Salerno	emanuele.salerno@isti.cnr.it
Ovidio Salvetti	Ovidio.Salvetti@isti.cnr.it

UCL, Uk

Fred Stentiford	f.stentiford@ee.ucl.ac.uk

TAU-VISUAL, Israel

Nahum Kiryati	nk@eng.tau.ac.il
Nir Sochen	sochen@math.tau.ac.il

UFR, Germany

Nikos Canterakis	canterakis@informatik.uni-freiburg.de

UPC, Spain

Miriam Leon	mleon@gps.tsc.upc.es
Joan Llach	Joan.Llach@thomson.net
Montse Pards	montse@gps.tsc.upc.es
Philippe Salembier	philippe@gps.tsc.upc.es

AUTH, Greece

Costas	Cotsaces cotsaces@aiia.csd.auth.gr
Irene Kotsia	ekotsia@aiia.csd.auth.gr
Nikos Nikolaidis	nikolaid@aiia.csd.auth.gr
Ioannis Pitas	pitas@aiia.csd.auth.gr
Vasilios Solachidis	vasilis@zeus.csd.auth.gr

ICCS-NTUA, Greece

Petros Maragos	maragos@cs.ntua.gr
Anastasia Sofou	sofou@cs.ntua.gr
Iasonas Kokkinos	jkokkin@cs.ntua.gr
Rapantzikos Konstantinos	rap@image.ntua.gr
Raouzaiou Amaryllis	araouz@image.ntua.gr
Ioannou Spyros	sivann@image.ntua.gr
Karpouzis Konstantinos	karpou@image.ntua.gr

INRIA-IMEDIA, France

Nozha Boujemaa	Nozha.Boujemaa@inria.fr
Valérie Gouet	Valerie.Gouet@inria.fr
Hichem Sahbi	Hichem.Sahbi@inria.fr
Anne Verroust	Anne.Verroust@inria.fr
Sabri Boughorbel	Sabri.Boughorbel@inria.fr

1 Introduction

This report provides an overview about recent methods and algorithms for processing video and image data. The report concentrates on techniques and methods that have been used in the broad context of multimedia understanding. It addresses the whole spectrum of required methologies from visual features (static and dynamic), organization of visual features, segmentation, saliency and attention, and object detection and recognition.

2 Visual Features for CBR

2.1 Local descriptors for content-based image retrieval

When considering sub-image retrieval or object recognition, local image characterization approaches provide better results than global characterization approaches classically based on color, texture and shape. It allows to gain robustness against occlusions and cluttering since only a local description of the patch of interest is involved. Such approaches are said to be *salient features-based*, because the information extracted from the image is condensed into limited but salient sites. Many kind of salient features can be envisaged: they can be represented by salient regions resulting from an image segmentation step [7, 55], or by non-connected image zones resulting from the construction of saliency maps as proposed by Itti *et al.* in [25]. They can also be represented by edges [13, 52], junctions, or special points [1]. This last case is the most encountered one and conducts to the extraction of points of interest (often called key points or salient points).Using points of interest is mainly motivated by observing that they provide the most compact representation of the image content by limiting the correlation and redundancy between the detected features. Indeed, contrary to an edge that exhibits a grey-level regularity along the contour curve, or a region in which all pixels fulfill some homogeneity criterion, there does not exist evident correlation nor dependency between non-connected and isolated salient points. Moreover, due to their sparse spatial distribution, object or image recognition based on salient points is much more robust to occlusions.

It is also important to note that the use of points of interest are often inspired from many psychovisual works [14, 38] that have shown that the sensitivity of the HVS (*Human Visual System*) is not uniformly distributed across the image content. This observation leads to saliency-based image indexing approaches [53] in which the first step consists in condensating the global information contained in the image into a limited number of feature values. Consequently, these salient features are to be selected with precision and in accordance with the properties of the HVS since the computation of the signature will uniquely depend on them. Ideally, salient features should be robust to geometric transforms, slight changes

of viewpoint and variations of imaging conditions. The robustness regarding to coding schemes is also of prime importance for indexing purpose as underlined in [26]. Finally, most of the global image content should be contained in the extracted salient features or in a close neighborhood.

Image retrieval based on local descriptors follows a classical computation flow: first a robust salient feature detector must be designed. Such a detector determines the location of salient points in the image, plus the support regions around them where the local descriptor is to be computed. Second, a rich and compact descriptor is computed for each salient point, by analyzing the image information within the support region exhibited. Finally, a similarity measure must be found to compare two salient point signatures.

This survey is organized as follows: in section 2.1.1, we remind and briefly describe the point detectors we have encountered in the literature of Computer Vision. We revisit the classical Harris and Stephens point detector before presenting derived approaches that involve different support regions according to the involved image transformations. Other different techniques are also listed. In section 2.1.2, we revisit the local descriptors usually associated to such points of interest and support regions.

2.1.1 Point of interest detection

In the context of content-based images retrieval, the extracted points must have the following characteristics: they represent single points located in image area where the information is considered as perceptually relevant. Ideally, the point detector should have a good repeatability, i.e. should be able to repeat the extracted points from an image to another whatever the photometric/geometric transforms involved (translation, rotation, changes of scale, of viewpoint, of illumination, etc).

Since the early work of Moravec [43] for stereo matching, many point extractors have been proposed in the literature of Computer Vision. Few comparison studies has been done for these approaches. See for example the ones of 2000 for grey value images [49] and color images [19].

The most popular one is probably the *Harris and Stephens detector* [21], which has been used first for stereo purposes and then for image retrieval. It is based on the computation of the local auto-correlation function M of the signal at each pixel location. Large eigenvalues of this function indicate large curvatures in the two directions, which denotes the presence of a salient point. The eigenvalues computation is replaced by computing local maxima of the function $Det(M) - k.Trace(M)^2$. A modified version of this detector has been proposed in [48] by improving the computation of the spatial derivatives with precise gaussian derivatives. This precise version allows to gain in repeatability, as demonstrated in [49]. The precise version of the Harris detector has been also extended to deal with color images in [42], where it gets a better repeatability [19].

To obtain robustness to changes of viewing conditions, point of interest detectors should be invariant to image transformations. We revisit the more recent works that have been done for some usual image transformations. The encountered approaches provide specific point of interest detectors associated to specific support regions.

Scale invariance

The Harris detector (and its precise or color version) is invariant to rotation but is very sensitive to changes in image scale. Recent works proposed derived or new detectors to achieve scale invariance. The problem of identifying an appropriate scale for feature detection has been studied by Lindeberg who has described it as a problem of selection of a characteristic scale [31,32]. From these considerations, several works on scale invariance have been proposed for local features. They are described in the next paragraphs.

In [33], Lindeberg has found a stable keypoint location in scale space by searching for 3D maxima of a function based on the Laplacian normalized with the scale. On the other hand, Lowe considered the local extrema in scale-space of Differences of Gaussian images [35]. Such points of interest are often called *DoG points*. As demonstrated by Lowe and evaluated later in [40], the DoG approach represents a close approximation of the Laplacian one, which is successfully compared to other functions (the Gradient and the standard Harris functions).

The paper [40] also presents an extension of the Harris precise detector. Here, the *Harris-Laplace detector* is introduced. It consists in extracting Harris points at a characteristic scale: each point is localized in 2D with the Harris function and then in scale-space where the Laplacian attains a maximum over scales. The authors demonstrate that this detector provides the best repeatability according to other scale-space detectors like the Laplacian one, the DoG one, the Gradient one and the standard Harris.

All these approaches provide points of interest that are invariant to rotation and scale changes. The selected scale determines the size of the support region to consider around the point (usually a uniform gaussian kernel) during the local description step.

Affine transformations

Some recent approaches have been proposed to adapt the support region to affine transformations (skew and stretch). In [2], Baumberg extracts interest points at several scales and then adapts the shape of the support region to the local image structure using an iterative procedure based on the second order moment matrix. The uniform gaussian kernel which usually defines the support region is replaced by an ellipsoidal one.

In [41], the scale and the location of the points are directly extracted in an

affine invariant way. The Harris-Laplace detector is extended to cope with such a transformation. The key idea is that the eigenvalues of the second order moment matrix computed in a point can be used to normalize the region according to affine transforms. The properties of the second order matrix were also explored in [46], but their goal was to obtain an affine invariant texture descriptor.

Other techniques : spatial-frequency approaches

Different approaches have been proposed to extract points on sharp region boundaries instead of on corners. They are usually based on wavelet salient features detection. In [34], the authors propose a multi-resolution approach where salient points are associated to highest wavelet coefficient values. Paper [5] presents a similar approach that is independent of the wavelet filter size and that does not favor any contour direction. This approach also proposes a method for automatically determining the optimal number of salient points to extract.

2.1.2 Local image description

The second step consists in designing salient signatures that describe the support regions associated to the extracted points. Here the main task is to bridge the gap between image semantics and pixels. As underlined in [29], a saliency-based image representation paradigm can be defined by observing that two different approaches can be used to describe an image from a limited number of points of interest: a *global salient approach* and a *local salient approach*. In the former case, the technique consists in extracting a unique signature by considering globally the information contained in the support regions. This method can be considered as a tradeoff between classical global approaches working on the entire image and purely saliency-based approaches. Indexing proposals belonging to this class of methods include [50, 51, 56]. In the latter case, the approach consists in considering independently the information contained in each support region, resulting in the computation of a local signature per salient point. In this document, we are mainly focusing on this second kind of technique.

The description is a function of the photometric and geometric information around the point and be the most compact possible. Many techniques have been developed, they depend on the considered application and more precisely on the image transformations involved. Roughly speaking, two kind of local signatures can be envisaged: the signatures coming from the global image indexing approaches (color histograms, Gabor jets, etc.) and the purely salient signatures that are dedicated to the use of a salient point detector.

The simplest local description that can be associated to a point of interest has been proposed for stereo matching purposes [57]: it is the vector of the pixels located in a window around the point. A cross-correlation measure serves as similarity measure between two points. But the high dimensionality of such

a descriptor makes it unapplicable for content-based images retrieval purposes. More compact point descriptors must be designed.

The differential invariants family

A set of image derivatives computed up to a given order approximates a point of interest neighborhood. Such a set is usually referred to *local jet* and the local description obtained is invariant to image translation. A stable estimation of the derivatives can be obtained by convolution with Gaussian derivatives. From the work of Koenderink [27] and Florack [9,10] on the properties of local derivatives, a lot of work has been done on differential descriptors. Basically, the components of the local jet can be combined to obtain invariance to image rotation, providing differential quantities which are invariant under SO(2) group of similitudes. Such an approach has been used in [15,48] for image retrieval purposes and in [42] for stereo matching purposes. When considering grey value images, they need to be computed up to order 3 [48], providing a set of 9 invariants. A generalization to color images has been proposed in [42]. The use of the color information allows to reduce the description to the first order invariants, providing 8 color invariants less sensitive to noise than high order derivatives [16].

Another technique, referred as *steerable filters* [11], consists in computing the derivatives of the local jet in a particular direction. For instance, steering derivatives in the direction of the gradient makes them invariant to rotation.

Some similar approaches of local description are based on *complex filters*. In [2] and [47] such filters are derived from the family $K(x, y, \theta) = f(x, y)e^{i\theta}$, where θ is the orientation. [2] uses Gaussian derivatives for $f(x, y)$ whereas in [47] a polynomial is applied. These filters differ from the Gaussian derivatives by a linear coordinates change in filter response space. When dealing with rotation invariance, it is necessary to consider the modulus of each response filter (the phase is sensitive to rotation).

All these descriptors are invariant to 2D translation and rotation. To adapt them to scale changes, a solution consists first in computing the derivatives by using a Gaussian kernel parameterized by the characteristic scale founded during the salient point detection (see section 2.1.1) and second in normalizing such derivatives by this scale. See for example [40] where steerable filters are used as local descriptors. Another solution consists in directly normalizing the support regions associated to the points computed in scale-space: all supports are mapped to a circular region of constant radius. As a result, descriptors computed on such supports become invariant to scale changes. The same technique can be applied for affine transformations by mapping to a circular window the ellipsoidal support associated to point of interest extracted under affine transformations. The invariants obtained from this normalized support are invariant to affine transformations. Such an approach is employed in [39, 41].

When considering illumination changes, two approaches are possible to nor-

malize the invariants to affine illumination changes [18, 20]: first, it is possible to normalize the support region. Second, the derivatives can be normalized by dividing them with the appropriate power of the gradient magnitude. Such a normalization eliminates the linear factor. The differentiation operation naturally eliminates the offset one.

The SIFT descriptors

Based on the DoG detector [35], Lowe has proposed the *Scale Invariant Feature Transform approach* (SIFT) for describing the local neighborhood of such points. Here, the local neighborhood of the salient point is described with multiple images that are orientation planes representing a number of gradient orientations. This descriptor can be divided by the square root of the sum of squared components to obtain illumination invariance. This approach provides robustness against localization errors and small geometric distortions. Its main drawback is the high dimension of the feature space involved (128 gradient orientations). A complete use of The SIFT approach has been presented in [36] for reliable point matching between different views of an object or scene.

A recent performance evaluation of local descriptors [39] has shown that the SIFT descriptor performs best. Steerable filters come second. As noticed by the authors of this evaluation, this signature can represent a good choice given the low dimensionality.

Other techniques : spatial-frequency approaches

All other approaches are generally picked in the global image indexing literature. Many techniques which describe the frequency content of the images have been developed. The well known Fourier transform decomposes the image content into basis functions. But this representation makes not explicit the spatial relations between points and the basis functions are infinite, therefore difficult to adapt to a local approach. Frequency approaches based on the Gabor transform [12] allow to take spatial relations between points into account, see for example [56]. Approaches based on wavelets have been also explored [50, 54]. These classes of techniques are usually employed in the context of texture classification. Finally, the classical use of the color information have also been proposed [5,50]. However, all the obtained signatures need to be high dimensional to give a precise signature.

Similarity measures

Except for the SIFT descriptor where the distance measure is L2 [35], the similarity between descriptors is usually computed with the Mahalanobis distance. The involved covariance matrix takes the different magnitudes, possible correlations and variability of the feature components into account. Point descriptors are

subject to different kinds of noises: in practice, they are sensitive to image acquisition (sensors and sampling errors), to numerical errors, to points of interest delocalization, etc.

These considerations show the importance of the similarity measure which must be carefully chosen for the considered descriptor to achieve best performances. An optimal similarity measure is directly related to the shape of their variability. When considering the Mahalanobis distance, this noise can be integrated in the similarity measure via the covariance matrix Λ. When a model of noise of the components cannot be specified, the way to estimate the covariance of the components comes down to different empiric solutions:

- Estimating Λ from the whole available data. This simple solution generates weights that are not discriminant, since representing a rough model of noise. Even so, this is the most common solution encountered to compare features with the Mahalanobis distance;
- Estimating Λ from points of interest whose local neighborhood is submitted to synthetic photometric and geometric transformations and perturbations that usually apply to images;
- Estimating Λ from training sequences of real images. Several points on different images with representative perturbations are tracked and a combination of the covariance matrices obtained can be used as the model of noise of point characterization. This solution has been adopted for example in [48] for image retrieval and in [17] for object tracking in video sequences.

Adding geometrical constraints between points of interest

Semi-local geometric constraints that consider the spatial relations between neighbor salient points of the same image can be added to enrich the local point description [3, 15, 48]. Obviously, they depend on the image transformations involved by the application.

2.2 Multiresolution and content-driven representations

2.2.1 Classical approaches for multiresolution representations

There are several ways to transform one representation of a given signal into another one. The most classical example is the Fourier transform, where a signal is decomposed into sinusoidal waves. Such a decomposition gives the intensity of the fluctuations (frequencies) in the signal which is often of great importance. However, due to the infinite extent of the sinusoidal functions, any local signal characteristics (i.e., an abrupt change in the signal) are spread over the entire representation, thus making them 'invisible'. This is a serious drawback since

singularities and irregular structures often carry the most important information in signals. For instance, in images, discontinuities in the intensity may provide the location of the object contours which are particularly meaningful for recognition purposes. For many other types of signals such as electro-cardiograms or radar signals, the interesting information is given by transients such as local extrema. Furthermore, such singularities usually occur with different location and localization (i.e., range, scale) in time and frequency. Consequently, transform methods that represent the signal at multiple scales are better suited for extracting information than methods that represent the signal at a single scale.

Over the last century, scientists in different fields struggled to overcome limitations of the Fourier transform and to build representations of signals that are able to adapt themselves to the nature of the signal. On the one hand, to 'pick up' the transients without giving up the frequency information, the signal should be decomposed over functions which are well localized in time (or space) and frequency. This leads to so-called *time-frequency representations*. On the other hand, since signal structure depends on the scale at which the signal is being perceived, it should be analyzed at different scales or levels of resolution. This results in so-called *multiresolution representations* which, besides a time parameter, also contain a scale parameter.

2.2.2 Multiresolution approaches

MR methods span a very broad array of concepts and approaches. In this report, we will mainly focus on pyramid and wavelet representations. There are, however, many other MR techniques such as quadtrees, multigrid methods and scale-space representations, to name a few. They have the 'multiresolution paradigm' in common, but apart from that they differ in many respects, both in theory and in practice.

Pyramids

Pyramids have been recognized early as an interesting tool for computer vision and image coding [4]. A classical pyramid scheme consists of three steps: (*i*) deriving a coarse approximation of an input image, (*ii*) predicting this image based on the coarse version, and (*iii*) taking their difference as the prediction error. This defines the analysis part. At synthesis, the prediction error is added back to the prediction from the coarse version, guaranteeing perfect reconstruction. Iteration of the analysis part over the coarse approximation yields a pyramid representation of the original image as an approximation image at the lowest resolution and a set of detail images at successive higher resolutions.

Wavelets

Wavelets [37] are functions that are well localized in time and frequency and that can be used to decompose a signal into different frequency bands with dif-

ferent time resolutions. This leads to the wavelet transform. Of particular interest is the discrete wavelet transform, which applies a two-channel filter bank (with downsampling) iteratively to the low-pass band (initially the original signal). The wavelet representation consists then of the low-pass band at the lowest resolution and the high-pass bands obtained at each step. This transform is invertible and non-redundant. As such, the corresponding decomposition differs from various other MR decompositions such as pyramids, which are redundant, and scale-spaces, which are non-invertible in general. Both aforementioned properties, i.e., invertibility and non-redundancy, turn the discrete wavelet transform into a highly efficient and applicable representation for a broad range of signal and image processing tasks such as denoising and, particularly, compression.

The wavelet decomposition (and reconstruction) of a discrete signal from a resolution to the next one is implemented by a two-channel perfect reconstruction filter bank such as in Fig. 1.

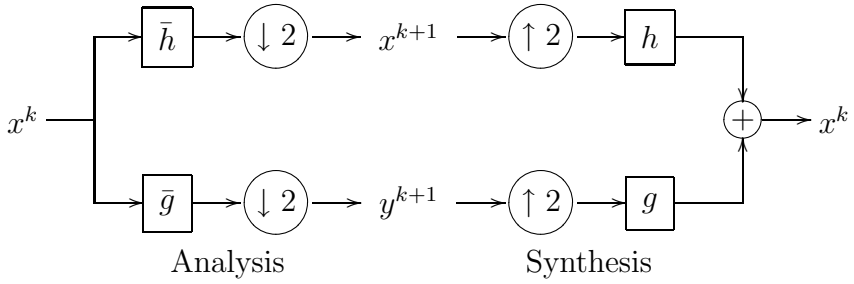


Figure 1: *Perfect reconstruction filter bank with low-pass and high-pass analysis filters $\bar{h}$, $\bar{g}$, respectively, and low-pass and high-pass synthesis filters h, g, respectively.*

There exists a great variety of wavelet families depending on the choice of the prototype wavelet or, alternatively, the filter's coefficients. However, imposing additional requirements such as orthogonality, symmetry, compactness of support, rapid decay and smoothness limits our choice. The 'optimal' choice of the wavelet basis will depend on the application at hand, and therein lies part of the difficulty of building a suitable wavelet representation.

Most wavelet (and pyramid) transforms have been designed in the one-dimensional case. By successive application of such one-dimensional transforms on the rows and the columns (or vice versa) of an image, one obtains a so-called *separable* two-dimensional transform. Non-separable transforms can also be constructed [22,28]. Although they provide decompositions with more general properties, they have been used less often in image applications due to the lack of general tools for their design.

2.2.3 The need for adaptive wavelets

Wavelets have had a tremendous impact on signal processing, both because of their unifying role and their success in several applications. The applicability

of the wavelet transform (as well as for other MR decompositions) is somewhat limited, however, by the linearity assumption. Coarsening a signal by means of linear operators may not be compatible with a natural coarsening of some signal attribute of interest (e.g., the shape of an object), and hence the use of linear procedures may be inconsistent in such applications. In general, linear filters smear the singularities of a signal and displaces their locations, causing undesirable effects.

Moreover, standard wavelets are often not suited for higher dimensional signals because they are not adapted to the 'geometry' of higher dimensional signal singularities. For example, an image comprises smooth regions separated by piecewise regular curves. Wavelets, however, are good at isolating the discontinuity across the curve, but they do not 'see' the smoothness along the curve. These observations indicate the need for MR representations which are data-dependent.

The importance of such 'data-driven' representations has led to a wealth of new directions in multiresolution approaches such as bandelets [30], ridgelets [8], curvelets [6], morphological wavelets [22], etc., which go beyond standard wavelet theory.

We will look for adaptive (i.e., content-driven) transforms which retain the desirable properties of the standard wavelet transform (e.g., non-redundancy and invertibility) while exploiting, in a simple way, the geometrical information of the underlying signal. This will allow for a better localization and representation of the singularities, as well as for sharper (perceptually better) approximations at lower resolutions.

In [23, 24, 44, 45] we have presented the construction of *adaptive wavelets* by means of an extension of the lifting scheme. The basic idea is to choose the update filters according to some decision criterion which depends on the local characteristics of the input signal. In this way, only homogeneous regions are smoothed while discontinuities are preserved. An interesting aspect of our approach is that it is neither causal nor redundant, i.e., it does not require any bookkeeping to enable perfect reconstruction. We show that these adaptive schemes yield lower entropies than schemes with fixed update filters, a property that is highly relevant in the context of *compression*. Despite all these attractive properties, a number of open theoretical and practical questions need to be addressed before such schemes become useful in signal processing and analysis applications. For example, we need to get a better understanding how to design update and prediction operators that lead to adaptive wavelet decompositions that satisfy properties key to a given application at hand. In our work, we have focused on binary decision maps and as a result, the adaptive scheme can only discriminate between two 'geometric events' (e.g., edge region or homogeneous region). In order to deal with the great richness of real-world signals and images, one must be able to incorporate the geometrical structure of the signals, for example, by using multiple criteria.

Another issue that needs to be addressed is the stability of the scheme. In particular, the behavior of the adaptive scheme under quantization needs a more

thorough investigation. Stability of decompositions is of utmost importance when they are being used in lossy compression schemes for image or video coding.

2.3 3D shape matching methods for content-based retrieval

A survey [ZP04] has been done recently on this subject and two papers have reported the result of comparative studies of the efficiency of several shape descriptors [SMKF04, TV04].

2.3.1 Specificity of 3D shapes

Shape representation In most of the 3D shape retrieval methods, the 3D shapes are represented by their boundary as a 2 dimensional surface such as a polyhedral surface, but they can also be voxelized.

Existing shape benchmarks (Princeton University [Dmse], Konstanz University [Dmsse] and Utrecht University [Dsre]) propose databases of polyhedral models but not give a guaranty on the way the surface is described (well defined or a soup of faces). Then some shape matching approaches will need a pre-processing phase to obtain a manifold surface.

Pose normalization The 3D models may have arbitrary scale and position. As most of the dissimilarity measures are sensitive to translations and rotations, it is necessary to put the models in a canonical coordinate system. Most of the approaches use a PCA transform (the method of [VSR01] is used in several approaches).
The PCA transforms do not give the axes orientation (as noticed in [OOIT02, ZP02, SKO00]) and may lead to ambiguities in the choice of the coordinate axes when the eigenvalues are similar. The dissimilarity measures of [ZP02, SKO00] take into account the coordinate systems obtained by switching the principal axes, which leads to a real invariance in rotation.

2.3.2 3D Shape matching methods

The approaches will be grouped into four categories:

- the methods based on a spatial decomposition of the 3D object,
- the methods working on the surface of the objects,
- the graph based methods,
- the 2D visual similarity based methods.

Spatial decomposition Most of the methods grouped in this section put the 3D shape in a spatial grid formed either by: voxels, concentric spheres, tetrahedra, etc... and use this spatial decomposition to compute the shape descriptor:
- **Voxels:**

in [PRM+00], a shape descriptor based on wavelets is proposed.

in [TV03], the descriptor is a set of weighted points, each weighted point being a "salient" point belonging to a voxel intersecting the surface. The dissimilarity measure uses the proportional transportation distance [GV02] derived from the Earth Mover's distance.

in [KCD+03], a reflective symmetry descriptor is built measuring the reflective symmetry of the object w.r.t. every plane passing by the centroid.
- **Subdivision of a sphere** [AKKS99]: the 3D object is put into a decomposition of a sphere in angular sectors and concentric spheres. A histogram based on this decomposition is built and a quadratic form distance measure is used to compare two objects.
- **Concentric spheres:** two approaches are based on spherical harmonics [VS02], [KF02]. Another is based on Zernike descriptors [NK03, NK04] and seems to obtain better results.
- **Tetrahedra** [ZC01]: the descriptor is computed using Fourier transforms on an approximation of the 3D object into a set of tetrahedra.
- **2D slices** [NK01]: the similarity is evaluated by computing 2D distance fields between the successive slices of the voxelized shapes and transforming them into a 3D distance between the two objects. This computation must be made for each couple of 3D objects and then it can be long.
- **principal axes of inertia** [OOIT02]: the 3D models are parameterized along the principal axes of inertia and three histograms (moment of inertia about the axis, average distance of the surface about the axis and variance of the distance from the surface to the axis) are computed.

Working on the surface of the objects These two approaches work only on well defined surfaces:
- **Shape index histogram (SF3D)** [ZP02]: a histogram of the shape index (function of the two principal curvatures on continuous surfaces) is adapted for polyhedral surfaces.
- **Curvature map** [ABP03]: the shape is represented by its curvature map, transforming the 3D problem into an image indexing problem. This method only works on genus 0 surfaces.

The other approaches work on the facets and do not need to have a manifold surface.
- **Complex EGI** [KI93]: a histogram is built on the Gaussian sphere. In this representation, the weight associated with each outward surface normal depends

on the facet's area and on the distance from the origin. This method is sensitive to change in facets orientations.
- **Moment-based** [PRM+00,ETA01]: these two approaches calculate the 3D statistical moments and compare them. The second approach adapts the similarity computation with SVM to obtain an interactive method.
- **Cord-based** [PRM+00]: three histograms describing the distribution of angles between the cords and the first (respectively the second) principal axis and the distribution of the length of the cords are built. The Hamming distance is used when comparing the histograms.
- **Shape distribution** [OFCD02, ILSR02] : these approaches compute a shape descriptor by combining probability distributions on shape functions representing the geometric properties of the 3D shape (angle between three points of the surface, distance between the centroid and a point, distance between two points of the surface, etc...). These methods achieve scale and affine invariance.
- **Hough transform** [ZP02]: a shape descriptor based on a 3D Hough transform is computed. The ambiguities of the PCA algorithm are taken into account in their method to obtain a rotation invariant shape descriptor.

Graph based approaches The main point of the graph based approaches is to extract the the topological information which may be lost in the other approaches. These methods are affine invariant but most of them only work on well defined polyhedral models. The similarity measure between shapes consists in a recursive graph comparison.
The method of [SSGD03] works on a voxelization of the shape and use a thinning algorithm to compute the skeletal nodes.
The skeletal graphs of [HSKK01, BSRS03, TS04] use Reeb graphs based on the computation of a geodesic distance on the surface.

2D visual similarity-based methods When the polyhedral shape is ill-defined, the similarity can be computed comparing their appearance:
- in [CK01], each object is associated to 72 viewpoints computed by rotating the camera along an axis. The views are then structured in a shock graph. The comparison of the view of an object with the views of the database models uses a shock graph matching.
- in [ONT03], after a normalization process, they compute depth images from 42 viewpoints and compare the shape feature vectors using Zhang's Fourier descriptors [ZL02].
- in [CTSO03], one hundred orthogonal projections are encoded by Zernike moments and Fourier descriptors for retrieval.

References

[1] C. Bauckhage and C. Schmid. Evaluation of keypoint detectors. Technical report, INRIA, 1996.

[2] Adam Baumberg. Reliable feature matching across widely separated views. In IEEE *Conference on Computer Vision and Pattern Recognition*, pages 774–781, 2000.

[3] N. Boujemaa, F. Fleuret, V. Gouet, and H. Sahbi. Automatic textual annotation of video news based on semantic visual object extraction. In *IS&T/SPIE Conference on Storage and Retrieval Methods and Applications for Multimedia*, 2004.

[4] P. J. Burt and E. H. Adelson. The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, April 1983.

[5] C. Laurent, N. Laurent, and M. Visani. Color image retrieval based on wavelet salient features detection. In *Intl. Workshop on Content-Based Multimedia Indexing*, 2003.

[6] E. J. Candès. The curvelet transform for image denoising. In *Proceedings of the IEEE International Conference on Image Processing*, Thessaloniki, Greece, October 7-10 2001.

[7] Carson C., Belongie S., Greespan H., and Malik J. Blobworld: Image Segmentation Using Expectation-Maximization and Its Application to Image Querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1026–1038, August 2002.

[8] D. L. Donoho. Orthonormal ridgelets and linear singularities. Technical report, Statistics Department, Stanford University, California, 1998.

[9] L.M.J. Florack, B.M ter Haar Romeny, J.J. Koenderink, and M.A. Viergever. General intensity transformations and second order invariants. In *7th Scandinavian Conference on Image Analysis*, pages 338–345, Aalborg, Denmark, 1991.

[10] L.M.J. Florack, B.M ter Haar Romeny, J.J. Koenderink, and M.A. Viergever. General intensity transformations and differential invariants. *Journal of Mathematical Imaging and Vision*, 4(2):171–187, 1994.

[11] W.T. Freeman and E.H. Adelson. The design and use of steerable filters. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906, 1991.

[12] D. Gabor. Theory of communication. *Journal of the Inst. Elec. Eng.*, 93(26):429–457, 1946.

[13] Gevers T. and Smeulders W.M. PicToSeek: Combining Color and Shape Invariant Features for Image Retrieval. *IEEE Transactions on Image Processing*, 9(1):102–119, January 2000.

[14] Gordon I.E. *Theories of Visual Perception*. Wiley, second edition edition, 1997.

[15] V. Gouet and N. Boujemaa. Object-based queries using color points of interest. In *IEEE Workshop on Content-Based Access of Image and Video Libraries (CBAIVL 2001)*, pages 30–36, Kauai, Hawaii, USA, 2001.

[16] V. Gouet and N. Boujemaa. On the robustness of color points of interest for image retrieval. In IEEE *International Conference on Image Processing (ICIP'2002)*, Rochester, New York, USA, September 2002.

[17] V. Gouet and B. Lameyre. SAP: a robust approach to track objects in video streams with snakes and points. In *To appear in the British Machine Vision Conference*, Kingston University, London, UK, September 2004.

[18] V. Gouet and P. Montesinos. Normalisation des images en couleur face aux changements d'illumination. In *13me Congrs Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle*, volume II, pages 415–424, Angers, France, 2002.

[19] V. Gouet, P. Montesinos, R. Deriche, and D. Pelé. Evaluation de détecteurs de points d'intérêt pour la couleur. In *Reconnaissance des Formes et Intelligence Artificielle*, volume II, pages 257–266, Paris, France, 2000.

[20] P. Gros. Color illumination models for image matching and indexing. In *Proceedings of* 15^{th} *International Conference on Pattern Recognition*, pages 580–583, Barcelona, Spain, September 2000.

[21] C. Harris and M. Stephens. A combined corner and edge detector. *Proceedings of the* 4^{th} *Alvey Vision Conference*, pages 147–151, 1988.

[22] H. J. A. M. Heijmans and J. Goutsias. Nonlinear multiresolution signal decomposition schemes. Part II: Morphological wavelets. *IEEE Transactions on Image Processing*, 9(11):1897–1913, 2000.

[23] H. J. A. M. Heijmans, B. Pesquet-Popescu, and G. Piella. Building nonredundant adaptive wavelets by update lifting. Submitted to *Applied and Computational Harmonic Analysis*, preliminary work: Research Report PNA-R0212, CWI, Amsterdam, 2003.

[24] H. J. A. M. Heijmans, G. Piella, and B. Pesquet-Popescu. Adaptive wavelets for image compression using update lifting: Quantisation and error analysis. Submitted to *International Journal of Wavelets, Multiresolution and Information Processing* (IJWMIP), 2004.

[25] Itti L., Koch C., and Niebur E. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, November 1998.

[26] Jean-Michel Jolion and Stphane Bres. Influence du codage jpeg sur des descripteurs d'images. *Traitement du signal*, 15(4):309–320, 1999.

[27] J.J. Koenderink and A.J. Van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 55:367–375, 1987.

[28] J. Kovačević and M. Vetterli. Nonseparable multidimensional perfect reconstruction filter banks and wavelet bases for $\mathbb{R}^n$. *IEEE Transactions on Information Theory*, 38:533–555, 1992.

[29] Laurent C., Laurent N., Maurizot M., and Dorval T. In Depth Analysis and Evaluation of Saliency-based Color Image Indexing Methods using Wavelet Salient Features. *to appear in Multimedia Tools and Applications*, 2004.

[30] E. Le Pennec and S. G. Mallat. Image compression with geometrical wavelets. In *Proceedings of the IEEE International Conference on Image Processing*, Vancouver, Canada, September 10-13 2000.

[31] T. Lindeberg. On scale selection for differential operators. In *Proceedings of* 8^{th} *Scandinavian Conference on Image Analysis, Tromso, Norway*, pages 857–866, May 1993.

[32] T. Lindeberg. Scale-space theory: A basic tool for analysing structures at different scales. *Journal od Applied Statistics*, 21(2):224–270, 1994.

[33] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116, 1998.

[34] E. Loupias, N. Sebe, S. Bres, and J.M. Jolion. Wavelet-based salient points for image retrieval. In *IEEE Int. Conf. On Image Processing*, Vancouver, 2000.

[35] David G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, pages 1150–1157, Corfu, Greece, 1999.

[36] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Accepted for publication int the International Journal of Computer Vision*, 2004.

[37] S. G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:674–693, 1989.

[38] Marr D. *Vision*. W.H. Freeman and Company, 1982.

[39] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *Intl. Computer Vision and Pattern Recognition*, 2003.

[40] Krystian Mikolajczyk and Cordelia Schmid. Indexing based on scale invariant interest points. In *International Conference on Computer Vision*, Vancouver, Canada, July 2001.

[41] Krystian Mikolajczyk and Cordelia Schmid. An affine invariant interest point detector. In *European Conference on Computer Vision*, volume 1, pages 128–142, 2002.

[42] P. Montesinos, V. Gouet, and R. Deriche. Differential Invariants for Color Images. In *Proceedings of* 14^{th} *International Conference on Pattern Recognition*, Brisbane, Australia, 1998.

[43] H.P Moravec. Towards automatic visual obstacle avoidance. In *Proceedings of the* 5^{th} *International Joint Conference on Artificial Intelligence*, page 584, Cambridge, Massachusetts, Etats-Unis, August 1977.

[44] G. Piella and H. J. A. M. Heijmans. Adaptive lifting schemes with perfect reconstruction. *IEEE Transactions on Signal Processing*, 50(7):1620–1630, July 2002.

[45] G. Piella, B. Pesquet-Popescu, and H. J. A. M. Heijmans. Adaptive update lifting with a decision rule based on derivative filters. *IEEE Signal Processing Letters*, 9(10):329–332, October 2002.

[46] F. Schaffalitzky and A. Zisserman. Viewpoint invariant texture matching and wide baseline stereo. In *International Conference on Computer Vision*, pages 636–643, Vancouver, Canada, 2001.

[47] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets. In *European Conference on Computer Vision*, pages 414–431, 2002.

[48] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–534, May 1997.

[49] C. Schmid, R. Mohr, and Ch. Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2):151–172, 2000.

[50] N. Sebe, Q. Tian, Michael S. Lew, and T.S. Huang. Color indexing using wavelet-based salient points. In IEEE *Workshop on Content-based Access of Image and Video Libraries*, pages 15–19, 2000.

[51] Sebe N. and Lew M.S. Salient Points for Content-based Retrieval. In *Proc. of the British Machine Vision Conference*, Manchester-UK, 2001.

[52] Shim S. and Choi T. Edge Color Histogram for Image Retrieval. In *IEEE International Conference on Image Processing*, volume 3, pages 957–960, Rochester (NY), September 2002.

[53] Smeulders A.W.M., Worring M., Santini S., Gupta A., and Jain R. Content-Based Image Retrieval at the End of the Early Years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, December 2000.

[54] J.K.M. Vetterli. *Wavelets and Subband Coding*. Prentice Hall, 1995.

[55] Wang J.Z., Li J., and Wiederhold G. SIMPLicity: Semantics-Sensitive Integrated Matching for Picture Libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947–963, September 2001.

[56] Christian Wolf, Jean-Michel Jolion, Walter Kropatsch, and Horst Bishof. Content based image retrieval using interest points and texture features. In IAPR *International Conference on Pattern Recognition*, volume 4, pages 234–237, Barcelona, Spain, 2000.

[57] Zhengyou Zhang, Rachid Deriche, Olivier Faugeras, and Quang-Tuan Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence*, 78:87–119, December 1995.

References

[ABP03] J. Assfalg, A. Del Bimbo, and P. Pala. Retrieval of 3D objects using curvature maps and weighted walkthroughs. In *International Conference on Image Analysis and Processing (ICIAP 03)*, Mantova, Italy, September 2003.

[AKKS99] M. Ankerst, G. Kastenmller, H.-P. Kriegel, and T. Seidl. 3D shape histograms for similarity search and classification in spatial databases. In *6th International Symposium on Large Spatial Databases (SSD'99)*, volume 1651

of *Lecture Notes in Computer Science*, pages 207–226, Hong Kong, China, 1999. Springer.

[BSRS03] D. Bespalov, A. Shokoufandeh, W. C. Regli, and W. Sun. Scale-space representation of 3d models and topological matching. In *Proceedings of the eighth ACM symposium on Solid modeling and applications*, pages 208–215. ACM Press, 2003.

[CK01] C.M. Cyr and B.B. Kimia. 3D object recognition using shape similarity-based aspect graph. In *International Conference on Computer Vision (ICCV)*, pages 254–261, 2001.

[CTSO03] D.Y. Chen, X.P. Tian, Y.T. Shen, and M. Ouhyoung. On visual similarity based 3D model retrieval. *Computer graphics forum*, 22(3):223–232, September 2003.

[Dmse] Princeton 3D models search engine. http://shape.cs.princeton.edu/search.html.

[Dmsse] Konstanz 3D models similarity search engine. http://merkur01.inf.uni-konstanz.de/cccc/.

[Dsre] Utrecht 3D shape retrieval engine. http://www.cs.uu.nl/centers/give/imaging/3drecog/3dmatchi

[ETA01] M. Elad, A. Tal, and S. Ar. Content based retrieval of VRML objects - an iterative and interactive approach. In *EG Multimedia*, pages 97–108, September 2001.

[GV02] P. Giannopoulos and R.C. Veltkamp. A pseudo-metric for weighted point sets. In *European Conference on Computer Vision (ECCV 2002)*, pages 715–730. LNCS 2352, 2002.

[HSKK01] M. Hilaga, Y. Shinagawa, T. Kohmura, and T. L. Kunii. Topology matching for fully automatic similarity estimation of 3D shapes. In *SIGGRAPH 2001 Conference Proceedings*, pages 203–212., Los Angeles, August 2001. ACM SIGGRAPH, Addison Wesley.

[ILSR02] C. Yiu Ip, D. Lapadat, L. Sieger, and W. C. Regli. Using shape distributions to compare solid models. In *7th ACM Symposium on Solid Modeling and Applications*, Saarbrcken, June 2002.

[KCD+03] M. Kazhdan, B. Chazelle, D. Dobkin, T. Funkhouser, and S. Rusinkiewicz. A reflective symmetry descriptor for 3D models. *Algorithmica*, 38(2):201–225, November 2003.

[KI93] S. B. Kang and K. Ikeuchi. The complex EGI: a new representation for 3D pose determination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(7):707–721, July 1993.

[KF02] M. Kazhdan and T. Funkhouser. Harmonic 3D shape matching. In *SIGGRAPH 2002 Technical Sketch*, 2002.

[NK01] M. Novotni and R. Klein. A geometric approach to 3D object comparison. In *International Conference on Shape Modeling and Applications*, pages 167–175, Genova, May 2001.

[NK03] M. Novotni and R. Klein. 3D Zernike descriptors for content based shape retrieval. In *Proceedings of the eighth ACM symposium on Solid modeling and applications*, pages 216–225. ACM Press, 2003.

[NK04] M. Novotni and R. Klein. Shape retrieval using 3D Zernike descriptors. *Computer Aided Design*, 36(11):1047–1062, September 2004. Special issue on solid modeling theory and applications.

[OFCD02] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin. Shape distributions. *ACM Transactions on Graphics*, 21(4):807–832, October 2002.

[ONT03] R. Ohbuchi, M. Nakazawa, and T. Takei. Retrieving 3D shapes based on their appearance. In *5th ACM SIGMM Workshop on Multimedia Information Retrieval (MIR 2003)*, Berkeley, Californica, USA, November 2003.

[OOIT02] R. Ohbuchi, T. Otagiri, M. Ibato, and T. Takei. Shape-similarity search of three-dimensional models using parameterized statistics. In *Pacific Graphics 2002*, Beijing, October 2002.

[PRM+00] E. Paquet, M. Rioux, A. Murching, T. Naveen, and A. Tabatabai. Description of shape information for 2-D and 3-D objects. *Signal Processing : Image Communication*, 16:103–122, 2000.

[SKO00] M.T. Suzuki, T. Kato, and N. Otsu. A similarity retrieval of 3D polygonal models using rotation invariant shape descriptors. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC2000)*, pages 2946–2952, Nashville, Tennessee, October 2000.

[SMKF04] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser. The Princeton shape benchmark. In *Shape Modeling and Applications Conference, SMI'2004*, pages 167–178, Genova, Italy, June 2004. IEEE.

[SSGD03] H. Sundar, D. Silver, N. Gagvani, and S. Dickinson. Skeleton based shape matching and retrieval. In *Shape Modeling and Applications Conference, SMI 2003*, Seoul, Korea, May 2003. IEEE.

[TS04] T Tung and F. Schmitt. Augmented reeb graphs for content-based retrieval of 3D mesh models. In *Shape Modeling and Applications Conference, SMI 2004*, pages 157–166, Genova, Italy, June 2004. IEEE.

[TV03] J.W.H. Tangelder and R.C. Veltkamp. Polyhedral model retrieval using weighted point sets. In *Shape Modeling and Applications Conference, SMI 2003*, Seoul, Korea, May 2003. IEEE.

[TV04] J.W.H. Tangelder and R.C. Veltkamp. A survey of content based 3D shape retrieval methods. In *Shape Modeling and Applications Conference, SMI 2004*, pages 145–156, Genova, Italy, June 2004. IEEE.

[VS02] D. V. Vranic and D. Saupe. Description of 3D-shape using a complex function on the sphere. In *IEEE International Conference on Multimedia and Expo (ICME 2002)*, pages 177–180, Lausanne, August 2002.

[VSR01] D. Vranic, D. Saupe, and J. Richter. Tools for 3D-object retrieval: Karhune-Loeve transform and spherical harmonics. In *2001 Workshop Multimedia Signal Processing*, Cannes, France, October 2001.

[ZC01] C. Zhang and T. Chen. Efficient feature extraction for 2D/3D objects in mesh representation. In *ICIP 2001*, Thessaloniki, 2001.

[ZL02] D. S. Zhang and G. Lu. Shape-based image retrieval using generic Fourier descriptor. *Signal Processing: Image Communication*, 17(10):825–848, November 2002.

[ZP02] T. Zaharia and F. Prteux. Shape-based retrieval of 3D mesh models. In *2002 IEEE International Conference on Multimedia and Expo (ICME'2002)*, Lausanne, August 2002.

[ZP04] T. Zaharia and F. Prêteux. 3D versus 2D/3D shape descriptors: A comparative study. In *Proceedings SPIE Conference on Image Processing: Algorithms and Systems III - IS & T / SPIE Symposium on Electronic Imaging, Science and Technology '03, San Jose, CA*, volume 5298, January 2004.

3 Object Detection and Recognition

3.1 Brief review on Object Recognition

Object recognition is the field of automatic description and classification of patterns [DHS00]. The recognition task may be 1) supervised where the object class is known, and some *a priori* informations are given or 2) unsupervised where the object is assigned to unknown class. Interest in the area of object recognition has been renewed due to the emerging applications such as handwritten identification, human/face recognition, image databases browsing, etc. Major approaches are presented in the following.

3.1.1 Template Matching

Template matching is one of the earliest approaches to object recognition. It is also known by the name of model-based object recognition [Agg95, Lon98, CF01, PS00]. The object to be recognized is matched against the template by taking into account various poses (translation and rotation) and scale changes. The template and the similarity measure (correlation) are optimized based on the available training set. Template matching is computationally consuming, but the availability of faster processors has now made this approach more feasible. The rigid template matching mentioned above, while effective in some application domains, has a number of disadvantages. For instance, it would fail if images presents distortion due to some processing or viewpoint change. Deformable template models [Gre93, BK89] can be used for object recognition when the deformation cannot be directly modelled.

3.1.2 Statistical Approach

The goal of the statistical approach is to establish decision boundaries given a set of training images [HAMJ01]. This boundaries should split feature space such that different image categories occupy compact and disjoint regions in a d-dimensional feature space. The effectiveness of the representation space is determined by how well images from different classes can be separated. In the statistical decision theoretic approach, the decision boundaries are determined by the probability distributions of the images belonging to each class, which must either be specified or learned [DGL96, DHS00, Vap98]. One can also take a discriminant analysis-based approach to classification: First a parametric form of the decision boundary (e.g., linear or quadratic) is specified; then the best decision boundary of the specified form is found based on the classification of training images. Non parametric approaches based on maximizing the margin between the decision boundaries and the training samples leads to the SVM classifiers [Vap98, Bur98, CHV99].

3.1.3 Syntactic Approach

In many recognition tasks, hierarchical representation is more appropriate to adopt. The image is viewed as being composed of simple sub-images which are themselves built from yet simpler sub-images [Fu82, Pav77] called primitives. In syntactic object recognition, an analogy can be drawn between the structure of images and the syntax of a language. The images are viewed as sentences belonging to a language, primitives are viewed as the alphabet of the language, and the sentences are generated according to a grammar [FB86a, FB86b]. Thus, a large collection of complex images can be described by a small number of primitives and grammatical rules. The grammar for each image class must be inferred from the available training samples. The implementation of a syntactic approach, however, leads to many difficulties which primarily have to do with the segmentation of noisy images (to detect the primitives) and the inference of the grammar from training data.

3.1.4 Neural Networks

Neural networks can be viewed as parallel computing systems with a large number of simple processors having many interconnections [Tho96, Zha00, Sme95]. Neural networks have the ability to learn complex nonlinear input-output relationships [EPdRH02], use sequential training procedures, and adapt themselves to the data. In spite of the seemingly different underlying principles, most of the well known neural network models are implicitly equivalent or similar to classical statistical object recognition methods. In [Rip93] a discussion is given on this relationship between neural networks and statistical object recognition.

3.2 Appearance-based object detection and recognition

Computer vision was born with the aim at building machines that "can see" [EC01]. This program has a twofold implication, since it may be intended either as a tool to understand the underlying properties and mechanisms of human vision, or as the theoretical basis for a set of applications aiming at extracting high level perceptual information from the analysis of natural scenes. The first aspect is mailnly related to similar applications of artificial intelligence (in the sense of simulating intelligent behavior); the second aspect is related to emulating intelligent responses to external stimuli, and has been studying all the perceptual cues that are useful to extract structure information from image data. A distinction exists between low-level vision and high-level vision, based on the increasing structuredness of the specific cues treated as primary data for processing. Attempts have been made to integrate different processing levels and different visual cues.

An extended notion of image analysis and understanding, however, needs

extended concepts of "images" and "objects" to be introduced, which are not necessarily "natural": any n-dimensional map of a scalar or vector quantity is referred to as an image.

The list of different applications of object detection and recognition would be practically infinite. We can safely say that almost every technological area is affected by this discipline. Medical diagnosis [DNPS03], industrial nondestructive testing [CACBS00], remote sensing [LDZ00], optical character recognition [VBT02], virtual/ augmented reality [ZCHS03], automatic image and video indexing [LB02], visual databases [FPZ04], security/ surveillance systems [CLK00] are just a few examples. In particular, remote sensing and security issues are assuming an increasing importance in facing the present natural or human-made threats.

Computer vision subfields are now very mature disciplines, in the sense that they have sound foundations in different branches of mathematics, statistics and information theory. Another pointer to maturity is the very large number of groups and researchers involved in this area, and the wide range of applications, especially in such fields as face recognition [ZCPR03], where decades of basic and applied research have produced efficient commercial off-the-shelf systems.

Very robust techniques have been developed and applied successfully in many real-world applications. Also, this has been made possible by the increasing capabilities of general-purpose and specialized hardware (ultra-fast CPUs, digital signal processors, etc.), which brought many tasks to quasi-real-time performances. To exemplify, pattern recognition techniques are being increasigly employed in medical imaging. Pattern segmentation and recognition are the artificial vision tools that permit to build robust systems in face and fingerprint recognition, in security/surveillance applications, and optical character recognition in automatic text processing. Another emerging field where object recognition has a relevance is in protection of the intellectual property. Besides common practice applications, a number of real-world problems remain unsolved for the lack of robust solutions. Normally, the existing algorithms are effective in at least partially controlled environments (indoor, structured...), but their performance degrades in more general cases. Both new insights in image/scene perception theory and new computational paradigms are now being exploited to solve problems that lack efficient solutions.

Object detection and recognition, just as all image science, is a highly interdisciplinary area. Since its early days, the views of engineers, mathematicians, physicists, psychologists and computer scientists were put together to find solutions at different levels of imaging and vision problems. All the recent algorithmic advancements are now being used to improve current solutions or to find solutions to new problems. Besides the classical tools of functional artificial intelligence (neural networks and algorithms [CU93]), the fields of statistical image and signal processing [CA02] are gaining new importance, also due to the availability of new probabilistic models and large image databases. New algorithmic

approaches, such as nature-inspired computation [FO2] [T03], complex systems and chaos [BDT99], etc., are also used, often showing both promising results and new deep insights.

As mentioned above, just a few detection/recognition procedures work robustly when faced with an unstructured (and dynamically changing) environment. As an example, we mention again the case of remote sensing and surveillance, with all the related subfields and critical applications, such as pollution control, disaster management, anti-terrorist measures. On one hand, this depends on the environment [SN00], but not totally on that [EC01]. The extraction of visual information can also depend on the attitude of the agent [MP97]. This fact has already been recognized in biological visual systems, where the identification of the different classes in a scene depends on the pre-attentive or attentive mood and on the specific task being performed. Further theoretical developments and interaction among different disciplines are needed to attack this problem, whereas classifying and recognizing all the objects in a scene still remains impossible. To find empirically effective and robust applications on the basis of current theoretical developments is thus the first challenge for the near future. This is a nontrivial task when truly real-world problems are to be solved. One of the proposed solutions is to rely on active vision systems [BPPP98], which are able to enrich the data space by capturing suitably chosen additional images.

The basic cue to accomplish object detection, classification and recognition is the feature. A feature can be any attribute of an image in any relevant space, such as the original pixel space or whatever other representation space (Fourier, wavelet...). Any image can thus be described by means of a feature vector.

With "appearance-based vision", we mean an approach to detect, classify, and recognize objects on the basis of sets of 2D images, or views. The concept of appearance-based vision is opposed to the one of "geometry-based vision", which relies on matching geometric primitives or templates. In the case of detection and classification of objects within complicated backgrounds, the appearance-based approach has proved to achieve better results than the approaches based of feature or template matching [ZCPR03].

The "features" on which appearance-based vision has to rely are not predetermined primitives or templates (possibily invariant to particular transformations), but are implicitly defined from the sample images chosen, from which they are to be learned [FFP03] [FPZ03]. This implies that large sets of training images should be available. Typical tasks are identification (to determine which component in a database, if any, is represented in the input image), verification (to determine whether or not an input image actually corresponds to a particular object in a database) [ZCPR03], and detection [RH02] (to determine presence and location of one or more objects of interest in an image). The choice of the analysis strategy depends on the final goal.

For appearance-based vision, the pixel (scalar or vector) values can simply be assumed as image features. This results, in general, in a very high-dimensional

feature space, thus implying the need for dimensionality reduction [BPPP98] [MK01] [NS98]. In any case (whether pixel-based or not), appearance matching can be performed by global (holistic) or local methods. Among global mehods, we find, for example, histogram matching and eigenspace representation. The first approach exploits suitable distances between histograms to establish a similarity between an input pattern and the items in a database. The second approach projects the input image onto the subspace spanned by the training set, in order to find the minimum-distance element and be able to label the input pattern.

The relevant image space can be built, for example, through principal component analysis (PCA [CA02]), or linear discriminant analysis (LDA [BHK97] [MK01]). In the PCA approach, a basis of orthogonal "eigenimages" is found [KS90] [TP91], from which it is possible to reconstruct any element of the original dataset to a fixed level of accuracy. Projecting an input image onto this basis and finding the minimum-distance element of the original dataset amounts to classify (or recognize) the image. The effectiveness of this strategy is based on data variance, that is, the eigenimages computed are generalized vectors oriented in the same direction as the eigenvectors of the data covariance matrix. This is not always a valid criterion to discriminate between different classes. In the LDA approach (or *Fisher analysis*), indeed, the selection criterion is based on maximizing the between-class variances over the within-class variances, instead of maximizing the data variances.

Both PCA and LDA rely on second-order statistics to reduce the dimensionality of the image space. However, important information may be encoded in higher-order statistics, so that higher-order methods should outperform second-order methods [BMS02]. In other words, the common redundancy that results from sensed data is thought of as originated by the combination of independent components. This is the rationale of the independent component analysis (ICA) approaches. In [BMS02], two different ICA approaches and a hybrid approach are shown, with application to face recognition. One possibility is to derive a set of independent images on which the training data are projected, as is done by the PCA approach. In this case, the basis faces are not only uncorrelated, but independent of each other. Another possibility is to evaluate a number of components whose combinations can give the individual faces in the database via independent coefficients. In this way, a factorial code for face images is found.

Statistical data processing has also revealed its usefulness in tasks that are only indirectly related to object detection and recognition. As ICA is a typical strategy to perform blind source separation [HO00], it can be used to extract or remove interfering patterns from the raw data prior to proper recognition. This has been done, by ICA or similar techniques, in astrophysical image processing [KBPST03] and in document image analysis [TBS04]. In the former case, blind source separation is used as a sort of object detection tool, aiming at separating statistically distinct diffuse patterns superimposed to one another. In the latter, the aim is to reduce or cancel various degradations typically affecting ancient

documents, in order to improve the error rate of optical character recognition systems or the legibility by a human reader. This can be done by both ICA and different color decorrelation approaches [TSMB04], among which PCA, applied to color or multispectral images of the document pages.

When their basic assumptions are not verified, the ICA-based approaches can fail their goal. Research is being done to extend the conditions of separability in these cases. Moreover, flexible algorithms are being developed to possibly take prior information into account, when the problem at hand is not totally blind (for example, when the statistical distribution of the sources is known, or when prior knowledge on the data model is available).

In particular, there is an additional hypothesis besides mutual independence that requires nongaussian distributed sources (except at most one). It has been shown that, for nonstationary processes, separation can be achieved even if more than one sources are Gaussian. If the sources are known to possess temporal structures, then the independence assumption can be dropped. The resulting separation techniques are commonly denoted as *dependent component analysis* (DCA) [BAR00] [BC03] [BBBBH03] [DCP03].

Another hypothesis that does not need to be verified is the linearity of the source mixture. Especially when there is a sensible within-class variability (see [BMS02] for face recognition), this hypothesis cannot be considered true. To overcome this difficulty is not an easy task, and no general solution has been found so far. Research is active in determining the kind of nonlinearities that allow the problem to be efficiently solved [CA02] [YAC98] [BJ02].

An crucial problem when working recognition systems are to be developed is benchmarking. Indeed, some standardized procedure to evaluate different products is necessary, along with a common test database, in order to allow a significant comparison to be made. To this end, very large image databases must be available, whose entries should be as similar as possible to the actual test images. This means that evaluation procedures and databases should be specifically application-oriented [ZCPR03]. Some work has been done for optical character recognition and fingerprint classification [BBT02] [BCGCW94], and for face recognition, where, among others, the different releases of the FERET face database and evaluation procedure are available [PWJR98] [PMRR00]. Benchmarking is particularly important in the mentioned applications, where many commercial off-the-shelf systems are now available. Refining the evaluation procedures would further help monitoring the performances of the different algorithms in their evolutions, and identifying the most urgent and/or promising research areas. At present, for example [ZCPR03], two major problems in face recognition are recognition under illumination and pose variations.

3.3 Face detection

Many methods for face detection are discussed in the literature, including artificial neural networks [RBK98] [Sun96], support vector machines [OFG97] [EPPP00] [RTSB01], Bayesian inference [CWT00], deformable templates [MYW+99], graph-matching [LBP95], skin color learning [HAMJ01] [SB00] and coarse-to-fine processing [FG01] [VJ01]. Distinguishing factors include whether they can solve the face detection problem with real complex backgrounds and the run-time cost.

3.3.1 Finding faces on simple background

Color provides a computationally efficient method which is robust under rotations in depth, partial occlusions and can be used to model and filter skin color efficiently from a training set using standard classifiers [SB02, HAMJ01, YLW98, CG99]. Face detection can also be achieved using motion. The general idea is to detect differences between the current and the previous frame in a video sequence. If the difference between pixel values is greater than a given threshold, the movement is considered to be significant and the pixel is set to be of interest. Many prior knowledge (cf. below) can be used to decide whether a pixel of interest belong to a face or not [EJ95, ST98].

It is also advantageous to use prior knowledge. For face detection, we know that the head is located at the top of the body, that a human normally walks upright, etc [**?**, YC98]. For example, blinking patterns in an image sequence is an easy and a reliable mean to detect the presence of a face since blinking provides a space-time signal which is easily detected and unique to faces [GBGB01]. The fact that both eyes blink together provides a redundancy which makes it possible to discriminate faces from other motion in the scene. Furthermore, symmetry and the fixed separation between the eyes provides a way to estimate the size and the orientation of the head.

3.3.2 Finding faces in a complex background

This is the most interesting, challenging, and practical case, since it serves many applications. The generic approach is based on modeling the face appearance using an a priori geometric or learned model. Many methods have been proposed to perform face detection in a complex background using machine learning techniques such as neural network classifiers [RBK98] [SP98], Bayesian inference [CWT00], support vector machines [OFG97] [RTSB01] and eigenfaces [MP95]. Other techniques are based on the analysis of facial structures, and include: graph matching [LBP95], geometrical hashing [LSW98], edge counting and coarse-to-fine processing [FG01], together with AdaBoost [VJ01] and FloatBoost [LZZ+02]. All these methods operate by extracting windows at different locations and scales; pre-processing subimages using normalization techniques and encoding them using an appropriate structure or a feature space. The

underlying extracted information is classified as face/non face using a suitable classifier and a search strategy. Again, techniques differ in the training model used, the amount of training data necessary to capture the complexity of the decision boundary, the facial representation, and mainly the search strategy used to reduce computation.

3.4 Face recognition

Public transactions based on passwords, cards, etc., show only that the information provided by the user is valid, not that their rightful owner is present. Face recognition is *a science of identification*, which substitutes a key or a password by the facial characteristics which are unique for a given individual. This science measures the statistical variability of faces using a human population and requires a preliminary step of face localization.

At this time it is not yet clear whether face recognition is *holistic* or a *local task*, i.e., if it depends on the whole face characteristics or some of them. In the human visual cortex, face recognition is a dedicated task different from object recognition as a dedicated part in the brain is responsible for this task. The psychophysics aspects have been largely studied by biologists in order to understand face recognition in the human visual cortex [YE89] [BHB93] and this will help developing original and effective face recognition algorithms.

In the remainder of this abstract, we will review the main representative state of the art techniques for face recognition based on modeling the face appearance. Existing techniques can be classified either as local or holistic [BP93, LLN00], they use intensity or infrared images [WPJW96] [SWNE01], their query mode can be face images or sketches [UL96] [kon96], and the nature of the task can be identification or verification [ZCR00]. We will illustrate the methods from the holistic or local processing point-of-views.

3.4.1 Holistic versus local methods

Holistic methods consider the whole face image in the recognition process. The basic method is correlation [BP93] which declares two faces as similar if their inner product is relatively high. This method is simple but sensitive to changes in the viewing conditions and the lighting effects. More sophisticated face recognition methods have been introduced among them the well studied Eigenfaces and Fisherfaces [SK87, MN95, PMS94, TP91]. The eigenface method finds the principal axes, in an Euclidean space, which maximize the variance of the faces through a training set. Then, faces are projected into a subspace spanned by the principal axes and the coefficients of projection will be used as a face description. In contrast to eigenface, fisherface finds the principal axes which make the between-class variance large while maintaining faces related to the same individual with a small within-class variance [BHK97]. Active appearance mod-

els [CWT00, LW99] combine both shape and texture characteristics in order to train a projection matrix which is used to infer the parameters of face identity, pose and shape variations. Other methods for face detection, based on statistics and geometry, have been introduced including optical flow [LCK02], neural networks [CF90, HCZZ00], support vector machines [JKYL00, GLC00, LGL00], Hidden Markov models [Rig01, SY94, Nyf95, AB96, Eic02] and linear subspaces [Sha92,OvdM02,BJ03,BH01]. The drawback of holistic methods resides mainly in their sensitivity to partial occlusion, variations of the contrast and the pose. Local methods have been introduced to overcome these drawbacks and they usually proceed by extracting some facial components (mouth, nose, etc.) [GHL71,KK72, CB65, Ble66] [LHLM02, RY92, Hal91, kan73] and by combining these components using appropriate classifiers [GLC00,HHP01,XPL02,HHP01,GZ01,TV00,CSS00].

Several methods for face component extraction and analysis have been introduced. A generalized symmetry operator is used in [RY92] to find the eyes and mouth locations. This approach stems from the almost symmetric nature of the face about a vertical line through the nose and a particular symmetry measure is maximized. [Hal91] used a template-based approach in order to locate the facial components. [kan73] introduced a technique based on the use of a Laplacian operator in order to obtain a binary image and the facial components. [BP93] used instead a gradient operator in order to derive their edge map, which is partitioned into vertical and horizontal image gradients. The horizontal image gradient is used to extract the left and the right boundaries of face and the nose while the vertical image gradient is used to detect the head-top, the eyes, the nose base and the mouth. Face outline detection was performed using dynamic programming principle as the outline of a face can be followed by a line, so the order can be defined without ambiguity. The issues of the accuracy of face component extraction has been addressed in [Zha99]. In many systems, good recognition results are dependent on accurate component extraction and registration, so the performance may degrade when these components are not determined accurately.

Malsburg introduced a face recognition algorithm based on graphs. A face is seen as a graph where its nodes correspond the facial components and the arc-labels are the distances between these components. The facial components are extracted by maximizing a correlation between a bunch of Gabor filters related to a prototype graph. Afterward, each node in the graph is moved locally in order to maximize further this correlation and to localize the underlying facial component with higher precision. Face recognition is based on graph matching. In [GHL71, KK72, CB65, Ble66], the overall configuration of a face can be represented by a vector of numerical data representing the relative position and the size of the main facial components and the face outline coordinates. In [LHLM02], the authors use an augmented Gabor feature vector as a concatenation of multiple responses of Gabor filters at different orientation, scales and locations.

Statistical methods have also been used to recognize faces using their local components among them support vector machines [GLC00, HHP01, XPL02] and

Ada-boost [GZ01]. For instance [HHP01] trained many SVM classifiers which are used as filters in order to locate the facial components (eyes, nose, tips of the nose, corners of the mouth, etc). The extracted components are used as input to a multi-class SVM which returns the identity of the aggregate face components. [GZ01] adapt the Ada-boost method [CSS00] for face recognition. Ada-boost is a method to combine a collection of weak classifiers (the weak learners are related to the facial components) to form a strong classifier after many iterations.

3.4.2 Hybrid methods

Hybrid methods and sensor fusion have received significant attention in the last decade. Sensor fusion has already been used in the combination of speech and face recognition for person identification. In the work of [**?**], the conclusion was that the future in person identification lies in hybrid recognition systems.

The work of [Gor95] was the first attempt to conceive such a hybrid recognition system combining two template based face recognition systems for both frontal and profile faces. [AB96] introduced a parallel hybrid recognition system based on three face classifiers, namely the profile approach by [YJB95], an HMM similar to the one in [SY94], and the EigenFace approach by [TP91].

3.5 Emotions detection

There is a long history of interest in the problem of recognizing emotion from facial expressions [EP78], and extensive studies on face perception during the last twenty years [P.73] [DM75] [SK84]. The salient issues in emotion recognition from faces are parallel in some respects to the issues associated with voices, but divergent in others.

As in speech, a long established tradition attempts to define the facial expression of emotion in terms of qualitative targets, i.e. static positions capable of being displayed in a still photograph. The still image usually captures the apex of the expression, i.e. the instant at which the indicators of emotion are most marked. More recently emphasis, has switched towards descriptions that emphasize gestures, i.e. significant movements of facial features.

In the context of faces, the task has almost always been to classify examples of archetypal emotions. That may well reflect the influence of Ekman and his colleagues, who have argued robustly that the facial expression of emotion is inherently categorical. More recently, morphing techniques have been used to probe states that are intermediate between archetypal expressions. They do reveal effects that are consistent with a degree of categorical structure in the domain of facial expression, but they are not particularly large, and there may be alternative ways of explaining them - notably by considering how category terms and facial parameters map onto activation-evaluation space [KK00].

Analysis of the emotional expression of a human face requires a number of preprocessing steps which attempt to detect or track the face, to locate characteristic facial regions such as eyes, mouth and nose on it, to extract and follow the movement of facial features, such as characteristic points in these regions, or model facial gestures using anatomic information about the face.

Facial features can be viewed [Ekm75] as either static (such as skin color), or slowly varying (such as permanent wrinkles), or rapidly varying (such as raising the eyebrows) with respect to time evolution. Detection of the position and shape of the mouth, eyes, particularly eyelids, wrinkles and extraction of features related to them are the targets of techniques applied to still images of humans. It has, however, been shown [Bas78], that facial expressions can be more accurately recognized from image sequences, than from a single still image. His experiments used point-light conditions, i.e. subjects viewed image sequences in which only white dots on a darkened surface of the face were visible. Expressions were recognized at above chance levels when based on image sequences, whereas only happiness and sadness were recognized at above chance levels when based on still images. Techniques which attempt to identify facial gestures for emotional expression characterization face the problems of locating or extracting the facial regions or features, computing the spatio-temporal motion of the face through optical flow estimation, and introducing geometric or physical muscle models describing the facial structure or gestures.

3.6 Text extraction methods in images and video sequences

he volume of data available nowadays in video format makes it necessary to create tools which allow to extract information from these video sequences in order to be classified (video indexing) or to be analyzed (video analysis) without human supervision. However, these tasks are very complex and remain an open problem. Therefore, the combination of different techniques could be of great interest to help solving this problem. Text extraction can be a key feature because it is usually synchronized and related to the scene in the sequence, obtaining extra information about the scene.

The contents in a video sequence can be perceptive, such as colour, shapes, textures, frequency changes, or semantic, such as objects or events and its relationships. The perceptive ones are easier to analyse automatically and the semantic ones are easier to handle linguistically. Therefore, since the computational cost of text extraction is lower than the cost of recognising objects, events or people, a good option is to detect and recognise text in a sequence.

At the moment there are many algorithms that are able to extract text from document files (OCR Optical Character Recognition) providing acceptable results. The main differences between OCR techniques and text recognition in video

sequences are that document files are usually binary images where, of course, letters are static and the background is white or black, while in video sequences background is complex and characters can move throughout time. OCR will be the last step in text recognition in video sequences.
Two kinds of text can be found in video sequences:

Scene text : It is composed by text that is integrated in the scene and is captured directly by the camera. It is more difficult to recognise because it can appear in any tilt or perspective, different illumination, on both planar and ridged surfaces, complete or partially hidden.

Caption text : It is the text that is introduced artificially over the frame. It can be static or in motion, but in any case its aim is to be readable and comprehensive by a viewer. In some references it is also called artificial text or graphic text.

Mainly all the different methods try to detect and recognize caption text, but some of them show also tools to deal with both texts. A minority is targeting only scene text. The aim of most methods is to find out the content in the video sequence, and their results are usually utilised for indexation. In this state of the art we are focusing on caption text methods.

3.6.1 Text features

Some of the main caption text features are the following.

Contrast between text and backgroud
Contrast is an important feature since in most images text must be readable, it cannot be blurred or occluded. Very often a high contrast is required as well as a steady brightness. One of the main problems localizing text is unsurprisingly the low contrast and the complex background. When characters have similar hue and brightness to the background their detection is almost impossible. In these cases, some enhancement tools must be employed as a pre-processing step in order to improve the contrast. Some standards for subtitling recommend using a black background. If coloured background is used then a legible text colour should be chosen. The most legible text colours on a black background are white, yellow, cyan and green, but it should be avoided using magenta, red and blue. Moreover, it has been checked empirically that for caption text there is no rule for choosing a colour. As a consequence, to avoid the possibility of characters blurring with background, they are highlighted by increasing the contour contrast.

Spatial cohesion

TYPOGRAPHY : Typography refers to the type of font used. Indeed, some of them are useful because their structure avoids confusions among the different

letters. This property takes into account not only able people, but physically handicapped too, like deaf viewers? This feature doesn't help directly for the localization but for character recognition system (OCR).

SIZE : An important fact is that text has to be readable. On one hand, researchers on vision have investigated human eye resolution, concluding that when letters are smaller than a minimal size, people cannot differentiate them. The minimal high and width sizes are approximately 15 and 7 respectively, but these values differ a little in each article. On the other hand, a maximal size is also bounded. Other important value is the ratio between height and width of a letter, which takes values around 0.9. Another size constraint is called word length or sentence length. Characters have similar heights and spacing within a text string. Therefore, this feature allows separating words.

COMPACTNESS (or FILLFACTOR) : If a bounding box containing the letter is build, compactness is the relation between pixels belonging to the letter and those belonging to the background. These features can be applied for both a simple character and an entire word. Allowed values used to be included in the interval [0.1, 1] [17], but it depends on the authors' criteria.

HORIZONTAL GEOMETRY : Text possesses certain orientation. To make text more easily readable it is usually displayed horizontally, in languages that are written horizontally.

Textured appearance :

The two previous features, contrast and spatial cohesion, can cause that the text search turns into texture segmentation. There are some papers which describe quite well what 'textured appearance' means, such us: "For example, by looking at the comic page of a news paper a few feet away, one can probably tell quickly where the text is without actually recognizing individual characters.", see [34]. Therefore, considering text as a whole entity, it has enough features to be detectable as a texture. Problems can appear when image textures and text features are very similar, like the leaves of a tree or grass in a field. Normally both textures and edges are fine structures and a level-pixel processing is required in order to localize and extract them.

Colour homogeneity :

Some papers take colour homogeneity as the main feature, because of the fact that colour segmentation preserves characters contour better than for example contrast segmentation, which may blur some edges [17]. Colour homogeneity could be analyzed in two different ways. On one hand, assuming that all the characters have the same colour, a whole word or line could be detected. On the other hand, each character might be written in different colour, but the letter colour itself remains homogenous. Both possibilities can be solved with the same algorithms. Summing up, monochrome characters are found more frequently than polychrome. Both of them can be detected with colour segmentation, but perhaps it could be assumed that polychrome characters are related to some artistic more than an informative purpose, so that some authors tend to discard this kind of

characters.

Strokes thickness :

Another feature that contributes to the textured appearance of the text is strokes thickness, because stroke is almost ever uniform. Thickness usually remains constant, except for some typography. In [24] the different features between Roman languages and non-roman languages can be found. One of them is precisely the stroke density: in non-roman languages density varies in the character itself, whereas it remains constant in roman languages. Another attribute related to stroke is its number in the character, which differs in both languages too. In Non-roman languages it fluctuates from 1 to 20, in Roman ones the variation is lower, from 1 to 4.

Temporal uniformity and redundancy :

People need time to read a sentence. This means that if every second 25 frames are displayed, the same caption text will be overlapped in so many frames as needed in order to make the sentence readable. Comprehension must be achieved by the reader at the time of viewing. Vision research has determined that humans need between two and three seconds in order to understand or process a complex image. Temporal uniformity affects not only the visualization time, but also the variation of the text size or their movement throughout the video sequence. These parameters can not sharply change from frame to frame. In this way this variation could be detected.

Mouvement on the frame :

This characteristic is related to the previous feature, but in this case it describes the most common text behaviour on the screen along the time:

- Static text. Characters present no movement, for example, when the name of a person appears on the screen, subtitles in a film, the scoreboard in a match, etc.

- Scrolling and crawling text. Normally it is linearly moving either horizontally from right to left (crawling), or vertically from bottom to top (scrolling). For example, the opening and closing credits on a film, last news, share prices information, etc. In roman languages to obtain the same level of comprehension, scrolling speed should be slower than crawling speed, because its movement is working against the reader's natural reading strategy. However, TV programs time is often limited and this additional information is not put on the screen to be fully readable.

- Flying text. This is the less common kind of movement. For example, it could be found in some TV advertisement and its movement is free, not predictable. As some papers stress, see [17], caption text aim doesn't fit with the effect achieved with flying text. In other words, flying text is not intended to give more information, but to attract attention. Its detection is also possible, but computational cost increases and the advantages of both

temporary uniformity and movement cannot be exploited. Moreover, this kind of text is more artistic and does not always have the same features, such as same character size or colour.

Due to the fact that flying text is quite improbable and usually not very interesting, when text candidates present neither static nor linear motion, candidate text blocks can be discarded. On the other hand, velocity is another discarding element. When candidate regions are moving faster than a threshold [pixels/frame], the regions are not designed in order to be read.

Position in the frame :
The main aim in this case is to avoid obscuring any important part or activity in the frame. Normally caption text is superimposed on the same area, which means caption text is normally found centred at the screen bottom. But it can be placed in a more appropriate place (e.g. football match scoreboard: top left/right corner). Sometimes, when the caption text in a frame is replaced for another one in a consecutive frame, the position of this new information differs a little with regard to the previous one.
In the literature, algorithms can be divided in two big groups, those that work in the compressed domain and those that work in the spatial domain. Many of them use the following parameters in order to evaluate its behaviour:

$Recall\ Recall = \frac{Correct}{(Correct+FalseNegatives)}$

$Precision\ Precision = \frac{Correct}{(Correct+FalsePositives)}$

where False Negatives are those characters, which have been discarded, and False Positives are those objects belonging to the background, which have been taken as characters.

3.6.2 Compressed domain techniques

In this group we include algorithms that are both in the compressed and in the semi-compressed domain.

Compressed domain : The only reference has been found in [2]. It is based on the localization of static characters over background in motion taking into account the macro-blocks belonging to P frames (MPEG-4). Moreover it assumes that text has horizontal geometry, that it does not occupy the whole frame and that it has to appear at least in three frames. These three features allow the algorithm to isolate macro-blocks and to determinate if the macro-blocks are candidates to contain text. Both recall and precision are high in those sequences with moving background and static text, like sports sequence (e.g. score in a football match). But it cannot be used in sequences containing moving text or static background.

Semi-compressed domain : This section is called semi-compressed domain, because algorithms don't work directly with macro-blocks but analysing the DCT (Discrete Cosinus Transformation) components [2], [4] and [40]. DCT plus motion compensation are utilised in the MPEG standard video compression in order to reduce spatial redundancy in a frame and temporal redundancy in consecutive frames, respectively. DCT coefficients represent spatial and directional periodicity. Thus, low level features can be directly extracted from compressed images. AC coefficients from horizontal harmonics show horizontal intensity variations; therefore, they would be high in case of having a text line. On the other hand, AC coefficients from vertical harmonics show vertical intensity variations; they would be high in case of having more than one single text line. In [26] a detailed explanation about the DCT coefficient meaning can be found. The DCT block size, the character size and their ratio are important. For instance, if each letter is bigger than the block size we will be evaluating a single letter stroke intensity variation, but not text intensity variation relative to the background. In the same way if the letter size is too small any texture could be analogous to text and easily confused (e.g. grass field). In [2] the algorithms are classified in edge-based method and correlation methods. The Edge-based methods take into account contrast between text and background, this is the reason why as a first step they calculate the Horizontal Intensity Variation in the DCT coefficients. The Correlation method [5] is applied only when a shot changes, so a shot detection must be previously done. In order to detect if the new shot contents text the intra-coded blocks increment is calculated in the B- and P-frames. Once the candidate blocks are chosen some text features are applied, such as that characters are made of strokes and its colour homogeneity and horizontal geometry, in order to localize text. In [40] this method is commented and discarded due to its vulnerability to scene changes.

Some other transformations could be used like the DWT (Discrete Wavelet Transform). This transformation gives more information than the previous one because spatial information is not lost with the transformation. Therefore, those areas with high values in high frequencies can be more easily found. Both [15] and [14] suggest wavelet transformation because of its capability to preserve spatial information. The text boxes are found through a hybrid: wavelet transformation and neural network. From the WT output some relevant statistical features can be chosen. In particular the more discriminant are the mean, the second and third order level calculated from the HL, LH and HH sub-bands. These vectors are used as input in a neural network.
In [3] some other transformations such as DHT (Discrete Haar Transform), DFT (Discrete Fourier Transform) and WHT (Walsh-Hadamard Transform), as well as the DWT (Discrete Wavelet Transform) are explained.
As a pre-processing tool, in [38] this kind of algorithms is used in a first step to

localize candidate areas.

3.6.3 Spatial domain techniques

Those methods that work with the pixel values and positions are called methods in the spatial domain and they can be classified according to the following image features:

- Edge-based [1],[2],[8] and [27]. Methods in this group are focused in the search of those areas that have a high contrast between text and background. In this way, edges from letters are identified and merged. Once these regions are recognised, spatial cohesion features are applied in order to discard false positives.

- Connected Components-based [16] and [18]. These methods use a bottom-up approach by grouping small components into successively larger components until all regions are identified in the image. Also in this case spatial cohesion features are applied.
  Both edge-based and Connected Components-based methods could be included under the same group, region-based methods.

- Texture-based [11],[13],[15],[34] and [35]. In this we can include many of the existing methods: they use the property that text in images have distinct textural properties that distinguish them from the background. Example are those which use the Gabor filter [11], Gaussian filter [35] or those based on the colour and shape of the regions [34] and [13]. If we had not classified into spatial and compressed domain, those methods based on wavelet or FFT would accomplish this textural properties.

- Correlation-based [33]. These methods can be summarized as those that use any kind of correlation in order to decide if a pixel belongs to a character or not.

- Others [29]. All the methods that have been mentioned don't use temporal information or use it as a complementary tool. In [29], temporal information is the main feature. After applying a shot detection technique, a vector through time for each pixel is calculated in a set of frames. The authors prove that, computing the PCA for each vector, feature vectors related to the background can be separated from those related to text. The main problem of this method is that it only can be applied when the sequence has static text and a moving background.

References

[1] Agnihotri,L.; Dimitrova,N.: 'Text Detection for Video Analysis', IEEE Workshop on CBAIVL, pp. 109-113, 1999

[2] Agnihotri,L.; Dimitrova,N.; Soletic,M.: 'Multi-layered Videotext Extraction Method', IEEE International Conference on Multimedia and Expo (ICME), August 26-29, 2002 Lausanne, Switzerland.

[3] Chaddha,N.; Gupta,A.:'Text Segmentation Using Linear Transforms', Proc. of Asilomar Conf. Circuits and Computers, pp. 422-427, 1996.

[4] Chen,D.; Odobez,J.-M.; Boulard,H.:'Text Detection and Recognition in Images and Video Frames', Pattern Recognition the journal of the pattern recognition society, accepted 20 June 2003.

[5] Gargi, U.; Antani, S.; Kasturi, R.: 'Indexing Text Events in Digital Video Databases', ICPR, pp. 916-918, August 1998.

[6] Gonzalez,R.C.: Woods, R.E.: 'Digital Image Processing', Addison-Wesley Publishing Company, Inc., 1992.

[7] Gu, L.; Kaneko,T.; Tanaka,N.: 'Morphological Segmentation Applied to Character Extraction from Color Cover Images', Mathematical Morphology and Its Applications to Image and Signal Processing, Kluwer Academic Publishers, ISMM'98, Amsterdam, The Netherlands, June 1998.

[8] Hua,X.-S., Chen, X.-R. et al.:'Automatic Location of Text in Video Frames', Intl Workshop on Multimedia Information Retrieval (MIR2001, In conjunction with ACM Multimedia 2001).

[9] Hua,X.-S., Yin, P., Zhang,H.-J.:'Efficient Video Text Recognition using multiplane frame integration', IEEE International Conference in Image Processing (ICIP 2002), Rochester, NY, USA, 22-25th September 2002.

[10] Hyvarinen, A.: 'Survey on Independent Component Analysis', "http://www.cis.hut.fi/ aapo/papers/NCS99web/node5.html", April 1999.

[11] Jain,A.K.; Bhatarcharjee,S.: 'Text Segmentation using Gabor filters for automatic document processing', Maching Vision and Application, 5:169 -184, 1992.

[12] Jain,A.K.; Yu,B.: 'Automatic Text Location in Images and Video Frames', Pattern recognition, Vol. 31, No. 12 pp. 2055-2076, 1998.

[13] Kim, H-K.: 'Efficient Automatic Text Location Method and Content-based Indexing and Structuring of Video Database', Journal of Visual Communication and Image Representation, Vol.7, No.4, pp. 336-344, December 1996.

[14] Li,H.; Doermann,D; Kia,O.: 'Automatic Text Detection and Tracking in Digital Video', Univ. of Maryland, College Park, Tech.Reps. LAMP-TR-028, CAR-TR-900 1998.

[15] Li,H.; Doermann,D; Kia,O.: 'Automatic Text Detection and Tracking in Digital Video', IEEE Trans. on Image Processing, Vol. 9, No.1, pp. 147-155, January 2000.

[16] Lienhart,R.; Stuber,F.:'Automatic Text Recognition in Digital Videos', Proceedings of SPIE Image and Video Processing IV 2666, pp. 180-188, 1996.

[17] Lienhart,R.; Effelsberg,W.:'Automatic Text Segmentation and Text recognition for Video Indexing', Technical Report TR-98-oo9, Praktische Informatik IV, University of Mannheim, May 1998.

[18] Lienhart,R.:'Text Segmentation and Text Recognition in Digital Videos', http://www.informatik.uni-mannheim.de/informatik/pi4/projects/-MoCA/Project-textSegmentationAndRecognition.html, MoCA Project, May 1998.

[19] Luo, B.; Tang,X.; Liu,J.; Zhang, H.:'Video Caption Detection and Extraction using Temporal Information', IEEE International Conference in Image Processing (ICIP 2003), Barcelona (Spain), September 14-17, 2003.

[20] Malik, J., and P. Perona.: 'Preattentive texture discrimination with early vision mechanisms'. Journal of the Optical Society of America A 7 Vol.5 pp. 923-931, 1990.

[21] Miene,A., Hermes, Th., Ioannidis, G.: 'Extracting textual Inserts from digital videos'. url: citeseer.nj.nec.com/miene01extracting.html, 2001.

[22] Myers,G.K.; Bolles,R.C.; Luong, Q.-T.; Herson, J.A.: 'Recognition of Text in 3-D Scenes', Fourth Symposium on Document Image Understanding Technology, Columbia, Maryland, April 2001.

[23] Qi,W., Gu,L., Jiang, H., Chen,X.-R., Zhang,H.-J.:'Integrating visual, audio and text analysis for news video', Image Processing, 2000. Proceedings 2000 International Conference on, pp. 520 -523 vol.3, 2000.

[24] Rosenfeld, Azriel; Doermann, David; DeMenthon, Daniel: 'Video Mining', Kluwer Academic Publishers (KAP), USA 2003.

[25] Sato,T.; Kanade,T.; Hughes,E.K.; Smith,M.A.; Satoh,S.: 'Video OCR: Indexing Digital News Libraries by Recognition of Superimposed Captions', ACM Multimedia Systems: Special Issue on Video Libraries, 1999. 7(5): pp. 385395. http://citeseer.nj.nec.com/sato99video.html

[26] Shen,B.; Sethi, I.K.: 'Direct feature extraction from compressed images', SPIE vol. 2670, Storage and Retrieval for Image and Video Databases IV, 1996.

[27] Smith,M.A.; Kanade,T.: 'Video Skimming and Characterization through the Combination and Language Understanding Techniques', IEEE Computer Vision and Pattern Recognition, pp. 775-781, 1997.

[28] Sun,J.; Yu,H.; Katsuyama,Y.: 'Effective Text Extraction and Recognition for www Images', Proceedings of the 2003 ACM symposium on Document engineering, Grenoble (France) pp. 115 - 117, November 20-22, 2003.

[29] Tang,X. et al.: 'Video Text Extraction using Temporal Feature Vectors', in Proc. of IEEE International Conference on Multimedia and Expo, Lausanne, Switzerland, August 2002.

[30] Tekinalp, S,; Alatan, A.A.:'Utilization of Texture, Contrast and Color Homogeneity for Detecting and Recognizing Text from Video Frames', IEEE International Conference in Image Processing (ICIP 2003), Barcelona, Spain, 14-17th September 2003.

[31] Wernicke,A., Lienhart,R.: 'On the Segmentation of Text in Videos', IEEE Int. Conference on Multimedia and Expo (ICME2000), Vol.3, pp. 1511-1514, July 2000; also Technical Report MRL-VIG00005, March 2000.

[32] Wong,E.K.; Chen,M.: 'A new Robust Algorithm for Video Text Extraction', Pattern Recognition 36 pp. 1397-1406 2003.

[33] Wong,E.K.; Chen,M.: 'A Robust Algorithm for Text Extraction in Color Video', Proceedings of IEEE International Conference on Multimedia and Expo, Vol. 2, pp. 797-800, 2000.

[34] Wu,V.; Manmatha, R.: 'TEXTFINDER: An Automatic System to Detect and Recognize Text in Frames', IEEE Transactions on Pattern analysis and machine intelligence, Vol. 21, No.11, pp. 1224 1229, November 1999.

[35] Wu,V.; Manmatha,R., Riseman,E.M.: 'Automatic text Detection and recognition', Proceedings of Image Understanding Workshop, pp. 707-712, 1997.

[36] Xi,J., Hua,X.-S., Chen,X.-R., Wenyin,L., Zhang.H.-J.: 'A Video Text Detection and Recognition System' 2001 IEEE International Conference on Multimedia and Expo (ICME2001), pp 1080-1083, August 22-25, Waseda University, Tokyo, Japan 2001.

[37] Zhang, Y., Chua,T.-S.: 'Detection of Text Captions in Compressed domain Video', Proceedings of ACM Multimedia'2000 Workshops (Multimedia Information Retrieval), California, USA. pp. 201-204, November 2000.

[38] Zhang, D., Rajendran, R. K., Chang, S.-F.:'General and Domain-Specific Techniques for Detecting and Recognizing Superimposed Text in Video', IEEE International Conference in Image Processing (ICIP 2002), Rochester, NY, USA, 22-25th September 2002.

[39] Zhong, David X.:'Color Space Analysis and Color Image Segmentation', VIP2000 Pan Sydney Area Workshop on Visual Information Processing, December 2000.

[40] Zhong,Y.; Zhang,H.; Jain,A.K.: 'Automatic Caption Localization in Compressed Video', IEEE Trans. PAMI, Vol. 22, No.4, pp. 385-393, April 2000.

References

[AB96] B. Achermann and H. Buncke. Combination of face classifiers for person indentification. *Proceedings of the 13th IAPR international conference on pattern recognition (ICPR)*, 3:416–420, 1996.

[Agg95] J. K. Aggarwal. A model-based object recognition in dense range image. *ACM Computing Survey*, 25(1):5–43, 1995.

[Bas78] J. Bassili. Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology*, 37:2049–2059, 1978.

[BH01] A. U. Batur and M. H. Hayes. Linear subspaces for illumination robust face recognition. *In proceedings of IEEE conference on computer vision and pattern recognition*, 2001.

[BHB93] V. Bruce, P. Hancock, and A. Burton. Human face perception and identification. *in face recognition: from theory to application (H. Wechsler, P.J. Phillips, V. Bruce, F.F. Soulie and T.S. Huang editors. Berlin Springer verlag*, pages 51–72, 1993.

[BHK97] P. Belhumeur, J. P. Hespanha, and D. Kriegman. Eigenfaces vs fisherfaces: Recognition using class specific linear projection. *IEEE transactions on pattern analysis and machine intelligence*, 19(7):711–720, 1997.

[BJ03] R. Basri and D. W. Jacobs. Lambertian reflectance and linear subspaces. *In Proceedings on pattern analysis and machine intelligence*, 25(2):218–233, 2003.

[BK89] R. Bajscy and S. Kovacic. Multiresolution elastic matching. *Computer Vision Graphics Image Processing*, 46:1–21, 1989.

[Ble66] W. Bledsoe. Man machine facial recognition. *Technical report Rep. PRI:22 Panoramic research Inc, Paolo Alto, Cal*, 1966.

[BP93] R. Brunelli and T. Poggio. Face recognition: Features versus templates. *In Pattern Analysis and Machine Intelligence.*, 15(10):1042–1052, 1993.

[BPNK97] H. J. P. BELHUMEUR P. N. and KRIEGMAN. Eigenfaces vs. fisherfaces: Recognition uing class specific linear projection. *IEEE Trans. Patt. Anal. Mach. Intell.*, 19:711–720, 1997.

[Bur98] C. Burges. A tutorial on support vector machines for pattern recognition. In *Data Mining and Knowledge Discovery*, volume 2, pages 121–167. 1998.

[CB65] H. Chan and W. W. Bledsoe. A man machine facial recognition system: some preliminary results. *Technical report, panoramic research, Inc, cal*, 1965.

[CF90] G. W. Cottrell and M. Fleming. Face recognition using unsupervised feature extraction. *In proceedings international Neural network conference*, 1:322–325, 1990.

[CF01] R. Campbell and P. Flynn. A survey of free-form object representation and recognition techniques. *Computer Vision and Image Understanding*, 81(2):166–210, 2001.

[CG99] J. Cai and A. Goshtasby. Detecting human faces in color images. *Image and Vision Computing.*, 18(1):63–75, 1999.

[CHV99] O. Chapelle, P. Haffner, and V. Vapnik. Svms for histogram-based image classification. *IEEE Transactions on Neural Networks*, 1999.

[CSS00] M. Collins, R. E. Schapire, and Y. Singer. Logistic regression, adaboost and bregman distances. In *Computational Learing Theory*, pages 158–169. 2000.

[CWT00] T. Cootes, K. Walker, and C. Taylor. View-based active appearance models. *In IEEE International Conference on Face and Gesture Recognition.*, pages 227–232, 2000.

[DGL96] L. Devroye, L. Gyrfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

[DHS00] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Hardcover, 2000.

[DM75] C. H. Davis M. *Recognition of Facial Expressions.New York: Arno Press*. 1975.

[Eic02] S. Eickeler. Face database retrieval using pseudo 2d hidden markov models. *In proceedings of IEEE conference on Face and Gesture*, 2002.

[EJ95] A. Eleftheriadis and A. Jacquin. Automatic face location and tracking for model assisted coding of video teleconferencing sequences at low bit rates. *Signal Processing: Image Communications.*, 7(3):231–248, 1995.

[Ekm75] F. W. Ekman, P. *Unmasking the Face. Prentice-Hall*. 1975.

[EP78] F. W. Ekman P. The facial action coding system. *Consulting Psychologists Press, San Francisco, CA.*, 1978.

[EPdRH02] M. Egmont-Petersen, D. de Ridder, and H. Handels. Image processing using neural networks - a review. *Pattern Recognition*, 35(10):279–301, 2002.

[EPPP00] T. Evgeniou, M. Pontil, C. Papageorgiou, and T. Poggio. Image representations for object detection using kernel classifiers. *In Asian Conference on Computer Vision*, pages 687–692, 2000.

[FB86a] K. Fu and T. Booth. Grammatical inference: Introduction and survey: Part i. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(3):343–360, 1986.

[FB86b] K. Fu and T. Booth. Grammatical inference: Introduction and survey: Part ii. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(3):360–376, 1986.

[FG01] F. Fleuret and D. Geman. Coarse-to-fine visual selection. *In International Journal of Computer Vision*, 41(2):85–107, 2001.

[Fu82] K. Fu. *Syntactical Pattern Recognition and Applications*. Prentice-Hall, 1982.

[GBGB01] K. Grauman, M. Betke, J. Gips, and G. Bradski. Communication via eye blinks detection and duration analysis in real time. *In Proceeding of Computer Vision and Pattern Recognition*, 2001.

[GHL71] A. J. Goldstein, L. Harmon, and A. B. Lesk. identification of human faces. *In proceeding IEEE*, 59:748, 1971.

[GLC00] G. Guo, S. Li, and K. Chan. Face recognition by support vector machines. *In proceedings on the international conference on Face and Gesture recognition*, pages 196–201, 2000.

[Gor95] G. Gordon. Face recognition from frontal and profil views. *Proceedings of the international workshop on automatic face and gesture recognition*, pages 47–52, 1995.

[Gre93] U. Grenander. General pattern theory. *Oxford University Press*, 1993.

[GZ01] G.-D. Guo and H.-J. Zhang. Boosting for face recognition. 2001.

[Hal91] P. Hallinan. Recognizing human eyes. *In spie proceedings: Geometric methods in computer vision*, 1570:214–226, 1991.

[HAMJ01] R. Hsu, M. Abdel-Mottaleb, and A. K. Jain. Face detection in color images. *In Proceedings of the IEEE International Conference on Image Processing*, pages 1046–1049, 2001.

[HCZZ00] F. J. Huang, T. Chen, Z. Zhou, and H.-J. Zhang. Pose invariant face recognition. *In proceedings of IEEE conference on Face and Gesture*, 2000.

[HHP01] B. Heisele, P. Ho, and T. Poggio. Face recognition with support vector machines : Global versus component-based approach. *In proceedings of International conference on computer vision*, 2001.

[JKYL00] K. Jonsson, J. Kittler, and J. M. Y.P. Li. Learning support vectors for face verification and recognition. *In proceedings of IEEE conference on Face and Gesture*, 2000.

[kan73] T. kanade. Picture processing by computer complex and recognition of human faces. *Technical report, kyoto university departement of computer science*, 1973.

[KK72] Y. Kaya and K. Kobayashi. A basic study on human face recognition. *In S. watanabe editors, frontiers of pattern recognition*, page 265, 1972.

[KK00] K. S. Karpouzis K., Tsapatsoulis N. Moving to continuous facial expression space using the mpeg-4 facial definition parameter (fdp) set. 2000.

[kon96] W. konen. Comparing facial line drawings with gray-level images: A case study on phantomas. *in proceedings, international conference on Artifical Neural neutworks*, pages 727–734, 1996.

[KS90] M. Kirby and L. Sirovich. Application of the karhunen-loeve procedure for the characterization of human faces. *IEEE Trans. Patt. Anal. Mach. Intell.*, 12, 1990.

[LBP95] T. Leung, M. Burl, and P. Perona. Finding faces in cluttered scenes using random labelled graph matching. *In Proceedings of the International Conference on Computer Vision*, pages 637–644, 1995.

[LCK02] X. Liu, T. Chen, and B. V. Kumar. On modeling variations for face authentication. *In proceedings of IEEE conference on Face and Gesture*, 2002.

[LGL00] Y. Li, S. Gong, and H. Liddell. Support vector regression and classification based multi-view face detection and recognition. *In proceedings of IEEE conference on Face and Gesture*, 2000.

[LHLM02] Q. Liu, R. Huang, H. Lu, and S. Ma. Face recognition using kernel based fisher discriminant analysis. *In proceedings of IEEE conference on Face and Gesture*, 2002.

[LLN00] R. Liao, S. Z. Li, and Nanyang. Face recognition based on multiple facial features. *In proceedings of IEEE conference on Face and Gesture*, 2000.

[Lon98] S. Loncaeic. A survey of shape analysis techniques. *Pattern Recognition*, 31(8):983–1001, 1998.

[LSW98] Y. Lamdan, J. Shwartz, and H. J. Wolfson. Object recognition by affine invariant matching. *In Proceedings of Computer Vision and Pattern Recognition*, pages 335–344, 1998.

[LW99] C. Liu and H. Wechsler. Face recognition using shape and texture. *In proceedings of IEEE conference on computer vision and pattern recognition*, 1999.

[LZZ+02] S. Li, L. Zhu, Z. Zhang, A. Blake, H. Zhang, and H.-Y. Shum. Statistical learning of multi-view face detection. *In Proceedings of the European Conference on Computer Vision.*, pages 67–81, 2002.

[MN95] H. Murase and S. K. Nayar. Visual learning and recognition of 3D objects from appearance. *ijcv*, 14(5):5–24, 1995.

[MP91] T. M. and PENTLAND. Eigenfaces for recognition. *J. Cogn. Neurosci*, 3:72–86, 1991.

[MP95] B. Moghaddam and A. Pentland. Probabilistic visual learning for object detection. *In Proceedings of the International Conference on Computer Vision*, pages 786–793, 1995.

[MYW+99] J. Miao, B. Yin, K. Wang, L. Shen, and X. Chen. A hierarchical multiscale and multiangle system for human face detection in complex background using gravity center template. *In Pattern Recognition*, 32(7):1237–1248, 1999.

[Nyf95] C. Nyffengger. Gesichterkennung mit hmm. *Masters thesis IAM*, 1995.

[OFG97] E. Osuna, R. Freund, and F. Girosi. Training support vector machines: an application to face detection. *In Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 130–136, 1997.

[OvdM02] K. Okada and C. von der Malsburg. Pose-invariant face recognition with parametric linear subspaces. *In proceedings of IEEE conference on Face and Gesture*, 2002.

[P.73] E. P. *Darwin and Facial Expressions. Academic Press.* 1973.

[Pav77] T. Pavlidis. *Structural Pattern Recognition.* Springer-Verlag, 1977.

[PMS94] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspace for face recognition. *iccv*, pages 84–91, 1994.

[PS00] R. Plamondon and S. Srihari. Online and off-line handwriting recognition: a comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):63–84, 2000.

[RBK98] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, 1998.

[Rig01] G. Rigoll. Hidden markov models for pattern recognition and man-machine-communication. *Pre-Conference Tutorial at 23. Annual Symposium for Pattern Recognition (DAGM 2001)*, 2001.

[Rip93] B. Ripley. *Statistical Aspects of Neural Networks*. U. Bornndorff-Neilsen, J. Jensen, and W. Kendal, eds., Chapman and Hall, 1993.

[RTSB01] S. Romdhani, P. Torr, B. Schlkopf, and A. Blake. Computationally efficient face detection. *iccv*, pages 695–700, 2001.

[RY92] D. Reisfeld and Y. Yeshurun. Robust detection of facial features by generalized symmetry. *In proceedings , International conference on pattern recognition*, pages 117–120, 1992.

[SB00] H. Sahbi and N. Boujemaa. From coarse-to-fine skin and face detection. *ACM International Conference on Multimedia.*, pages 432–434, 2000.

[SB02] H. Sahbi and N. Boujemaa. Robust face recognition using dynamic space warping. *In Proceedings of Springer Verlag Lecture Notes In Computer Science. ECCV's Workshop on Biometric Authentication.*, pages 121–132, 2002.

[Sha92] A. Shashua. Geometry and photometry in 3d visual recognition. *PhD thesis, Massachussets Institute of Technology*, 1992.

[SK84] E. P. Scherer K. *Approaches to Emotion, Lawrence Erlbaum Associates*. 1984.

[SK87] L. Sirovich and M. Kirby. Low dimensional procedure for the characterization of human faces. *In journal of opt soc. Am. A*, 4(3):519–524, 1987.

[Sme95] Y. Smetanin. Neural networks as systems for pattern recognition: a review. *Pattern Recognition and Image Analysis*, 5(2):254–293, 1995.

[SP98] K. K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):39–51, 1998.

[ST98] E. Saber and A. Tekalp. Frontal-view face detection and facial feature extraction using color, shape and symmetry based cost functions. *Pattern Recognition Letters*, 19(8):669–680, 1998.

[Sun96] K.-K. Sung. Learning and example selection for object and pattern detection,. *Ph.D. Thesis, Massachusetts Institute of Technology, Electrical Engineering and Computer Science*, 1996.

[SWNE01] D. A. Socolinsky, L. B. Wolff, J. D. Neuheisel, and C. K. Eveland. Illumination invariant face recognition using thermal infrared imagery. *In proceedings of IEEE conference on computer vision and pattern recognition*, 2001.

[SY94] F. Samaria and S. Young. Hmm based architecture for face identification. *Image and vision computing*, 12(8):537–543, 1994.

[Tho96] H. Thodberg. Review of bayesian neural networks with an application to near infrared spectroscopy. *IEEE Transactions on Neural Networks*, 7(1):56–72, 1996.

[TP91] M. Turk and A. Pentland. Eigenfaces for recognition. *In Journal of Cognitive Neuroscience*, 3(1):72–86, 1991.

[TV00] K. Tieu and P. Viola. Boosting image retrieval. *cvpr*, pages 228–235, 2000.

[UL96] R. Uhl and N. Lobo. A framework for recognizing a facial image from a police sketch. *In proceedings IEEE conference on computer vision and pattern recognition*, pages 586–593, 1996.

[Vap98] V. N. Vapnik. *Statistical learning theory*. J. Wiley and sons, 1998.

[VJ01] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *In Second International Workshop On Statistical and Computational Theories of Vision-Modeling, Learning, Computing and Sampling*, 2001.

[WPJW96] J. Wilder, P. J. Phillips, C. Jiang, and S. Wiener. Comparison of visible and infra-red imagery for face recognition. *Proceedings of the international conference on Automatic face and gesture recognition*, pages 182–187, 1996.

[WZR03] P. P. W. Zhao, R. Chellappa and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, 2003.

[XPL02] D. Xi, I. T. Podolak, and S.-W. Lee. Facial component extraction and face recognition with support vector machines. *In proceedings of IEEE conference on Face and Gesture*, 2002.

[YC98] K. C. Yow and R. Cipolla. Enhancing human face detection using motion and active contours. *accv*, pages 515–522, 1998.

[YE89] A. Young and H. Ellis. *Editors. Handbook of research on face processing. NORTH-HOLLAND*, 1989.

[YJB95] K. Yu, X. Jiang, and H. Bunke. Face recognition by facial profile analysis. *Proceedings of the international workshop on automatic face and gesture recognition*, pages 208–213, 1995.

[YLW98] J. Yang, W. Lu, and A. Waibel. Skin-color modeling and adaptation. *accv*, pages 687–694, 1998.

[ZCR00] W. Zhao, R. Chellappa, and A. Rosenfeld. Face recogntion: A litterature survey. *Technical report University of Maryland*, 2000.

[Zha99] W. Zhao. Improving the robustness of face recognition. *In proceedings of International conference on audio and video based person authentification*, pages 78–83, 1999.

[Zha00] G. Zhang. Neural networks for classification: a survey. *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews*, 30(4):451–462, 2000.

References

[BJ02] F. R. Bach, M. I. Jordan, "Kernel independent component analysis", J. Mach. Learn. Res., Vol. 3, 2002, pp. 1-48.

[BAR00] A. K. Barros, "Dependent component analysis", in M. Girolami, Ed., Advances in Independent Component Analysis, Springer Series Perspectives in Neural Computing, 2000, p. 63.

[BC03] A. K. Barros, A. Cichocki, "Extraction of specific signals with temporal structure", Neural Comput., Vol. 13, No. 9, September 2001, pp. 1995-2003.

[BMS02] M. S. Bartlett, J. R. Movellan, T. J. Sejnowski, "Face recognition by independent component analysis", IEEE Trans. on Neural Networks, Vol. 13, No. 6, November 2002, pp. 1450-1464.

[BBBBH03] L. Bedini, S. Bottini, C. Baccigalupi, P. Ballatore, D. Herranz, E.E. Kuruoğlu, E. Salerno, A. Tonazzini "A semi-blind approach for statistical source separation in astrophysical maps", ISTI-CNR, Pisa, Italy, Technical Report ISTI-2003-TR-35, October 2003.

[BHK97] P. N. Belhumeur, J. P. Hespanha, D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection", IEEE Trans. on PAMI, Vol. 19, No. 7, July 1997, pp. 711-720.

[BCGCW94] J. L. Blue, G. T. Candela, P. J. Grother, R. Chellappa, C. L. Wilson, "Evaluation of pattern classifiers for fingerprint and OCR applications", Pattern Recognition, Vol. 27, No. 4, April 1994, pp. 485-501.

[BDT99] E. Bonabeau, M. Dorigo and G. Theraulaz, Swarm Intelligence: From Natural to Artificial Systems, Oxford University Press, 1999.

[BPPP98] H. Borotschnig, L. Paletta, M. Prantl, A. Pinz, "Active Object Recognition in Parametric Eigenspace", Proc. Brit. Mach. Vis. Conf. 1998, Vol 2, pp. 629-638.

[BBT02] F.S. Brundick, A. E. M. Brodeen, M. S. Taylor, "A statistical approach to the generation of a database for evaluating OCR software", Int. J. on Docum. Anal. & Recogn., Vol. 4, No. 3, March 2002, pp. 170-176.

[CACBS00] G. Cavaccini, M. Agresti, M. Chimenti, E. Bozzi, O. Salvetti, " An evaluation approach to NDT ultrasound processes by wavelet transform", Nondestructive Testing, 2000, pp. 15-21.

[CU93] A. Cichocki and R. Unbehauen, Neural networks for optimization and signal processing, Wiley, 1993.

[CA02] A. Cichocki and S.I. Amari, Adaptive signal and image processing: Learning algorithms and applications, Wiley, 2002.

[CLK00] R. T. Collins, A. J. Lipton, T. Kanade, "Introduction to the special section on video surveillance", IEEE Trans. on PAMI, Vol. 22, No. 8, August 2000, pp. 745-746.

[DCP03] J. Delabrouille, J.-F. Cardoso, G. Patanchon, "Multidetector multicomponent spectral matching and applications for cosmic microwave background data analysis", Mon. Not. Roy. Astr. Soc., Vol. 346, No. 4, 21 December 2003, pp. 1089-1102.

[DNPS03] S. Di Bona, H. Niemann, G. Pieri, O. Salvetti, "Brain Volumes Characterisation Using Hierarchical Neural Networks", Artificial Intelligence in Medicine, 2003.

[EC01] J.-O. Eklundh and H. I. Christensen, "Computer vision: Past and future", in R. Wilhelm, Ed., Informatics: 10 years back, 10 years ahead, Lect. Not. Comp. Sci., Vol. 2000, 2001, pp. 328-340

[FFP03] L. Fei-Fei, R. Fergus, P. Perona, "A Bayesian Approach to Unsupervised One-Shot Learning of Object Categories", Proc. 9th Int. Conf. on Comp. Vis., Nice, France, October 2003, pp. 1134-1141.

[FPZ03] R. Fergus, P. Perona, A. Zisserman, "Object Class Recognition by Unsupervised Scale-Invariant Learning", Proc. IEEE Conf. Comp. Vis. Patt. Rec., October 2003, pp. 264-271.

[FPZ04] R. Fergus, P. Perona, A. Zisserman, "A Visual Category Filter for Google Images", Proc. 8th Europ. Conf. Comp. Vis., May 2004.

[FO2] A.A. Freitas, Data Mining and Knowledge Discovery with Evolutionary Algorithms, Springer Verlag, 2002.

[HO00] A. Hyvärinen, E. Oja, "Independent component analysis: algorithms and applications", Neural Networks, Vol. 13, 2000, pp. 411-430.

[KS90] M. Kirby, L. Sirovich, "Application of the Karhunen-Loeve procedure for the characterization of human faces", IEEE Trans. on PAMI, Vol. 12, No. 1, January 1990, pp. 103-108.

[KBPST03] E.E. Kuruoğlu, L. Bedini, M.T. Paratore, E. Salerno, A. Tonazzini, "Source separation in astrophysical maps using independent factor analysis", Neural Networks, Vol. 16, No. 3-4, April-May 2003, pp. 479-491.

[LB02] B. Le Saux, N. Boujemaa, "Unsupervised Categorization for Image Database Overview" VISUAL 2002, Lecture Notes in Computer Science, Vol. 2314, pp. 163-174.

[LDZ00] A. Lorette, X. Descombes, J. Zerubia, "Texture analysis through a Markovian modelling and fuzzy classification: Application to urban area Extraction from Satellite Images", Int. J. Comp. Vis., Vol. 36, No. 3, 2000, pp. 221-236.

[MK01] A. M. Martínez, A. C. Kak, "PCA versus LDA", IEEE Trans. on PAMI, Vol. 23, No. 2, February 2001, pp. 228-233.

[MP97] B. Moghaddam, A. Pentland, "Probabilistic Visual Learning for Object Representation", IEEE Trans. on PAMI, Vol. 19, No. 7, July 1997, pp. 696-710.

[NS98] R. Nelson, A. Selinger, "A cubist approach to object recognition", Int. Conf. Comp. Vis., Bombay, India, January 1998, pp. 614-621.

[PWJR98] P. J. Phillips, H. Wechsler, J. Juang, P. J. Rauss, "The feret database and evaluation procedure for face-recognition algorithms", Image Vision Computing J., Vol. 16, No. 5, 1998, pp. 295-306.

[PMRR00] P. J. Phillips, H. Moon, S. A. Rizvi, P. J. Rauss, "The FERET evaluation methodology for face recognition algorithms", IEEE Trans. on PAMI, Vol. 22, No. 10, October 2000, pp. 1090-1104.

[RH02] C. Rosenberg, M. Hebert, "Training Object Detection Models with Weakly Labeled Data", Proc. Brit. Mach. Vis. Conf. 2002, pp. 577-586.

[SN00] A. Selinger, R. C. Nelson, "Improving Appearance-Based Object Recognition in Cluttered Background", Proc. Int. Conf. Patt. Rec., Barcelona, Spain, September 2000, pp. 1-8.

[T03] R. Tateson et al., 2003, "Nature-inspired Computation: Towards Novel and Radical Computing", BT Technol. J., Vol. 18, No. 1.

[TSMB04] A. Tonazzini, E. Salerno, M. Mochi, L. Bedini, "Bleed-through removal from degraded documents using a color decorrelation method", ISTI-CNR, Pisa, Technical Report ISTI-2004-TR-04, February 2004.

[TBS04] A. Tonazzini, L. Bedini, E. Salerno, "Independent component analysis for document restoration", Int. J. on Docum. Anal. & Recogn., to appear, 2004.

[TP91] M. Turk, A. Pentland, "Eigenfaces for recognition", J. Cogn. Neurosci., Vol. 3, 1991, pp. 72-86.

[VBT02] S. Vezzosi, L. Bedini, A. Tonazzini, "An integrated system for the analysis and the recognition of characters in ancient documents", in Lecture Notes in Computer Science, Vol. 2423, 2002, pp. 49-52.

[YAC98] H. H. Yang, S. I. Amari, A. Cichocki, "Information-theoretic approach to blind separation of sources in nonlinear mixture", Signal Processing, Vol. 64, No. 3, 1998, pp. 291-300.

[ZCHS03] L. Zhang, B. Curless, A. Hertzmann, S. M. Seitz, "Shape and Motion under Varying Illumination: Unifying Structure from Motion, Photometric Stereo, and Multi-view Stereo", 9th IEEE Int. Conf. Comp. Vis., Nice, France, October 2003.

[ZCPR03] W. Zhao, R. Chellappa, P. J. Phillips, A. Rosenfeld, "Face recognition: A literature survey", ACM Computing Surveys, Vol. 35, No. 4, December 2003, pp. 399-458.

4 Segmentation

Image segmentation can be described as the process of partitioning the image into disjoint regions, each one being homogeneous and connected with respect to some cue(s), such as image intensity, texture, color, etc. Image segmentation methods can be classified as *boundary-based*, which generate an edge image which delineates the segments of the image, or *region-based*, which group image pixels based on the homogeneity of spatially localized features.

The cues that lead the segmentation process include most commonly image intensity, and usually additional features, like color and texture features, or motion and depth estimates. The incorporation of multiple-cue information in a segmentation process increases the potential of producing better results, when the corresponding features are treated appropriately.

Computer vision researchers have used different methodologies to attack the segmentation problem. The discrimination of different methodologies will be done according to the mathematical methodology they employ; the segmentation categories that will be presented are the following:

- Variational methods
- Statistical methods
- Graph-based methods
- Morphological methods

In the rest of this report, we will refer to recent advances in these four domains and present how they deal with multiple cues.

4.1 Variational Methods

In the variational framework the solution to the segmentation problem is expressed as the minimization of some energy functional J defined on partitions $\mathcal{R} = R_{1...N}$ of the image domain Ω; J is commonly expressed in terms of the contours ∂R_i and the interiors R_i of the image segments:

$$J(\mathcal{R}) = \sum_{i=1}^{N} \int_{R_i} f(x, R_i)dx + \sum_{i=1}^{N} \oint_{\partial R_i} g(s)ds \tag{1}$$

In the above formula f quantifies the homogeneity of the features at location x with those of region R_i and $g(s)$ is a decreasing function of edge strength.

Active Contours/Snakes: In the setting of Active Contours a closed curve or set of curves enclosing some initial region(s) is given, and we wish to modify them until they come into alignment with the contours of prominent image objects. This is accomplished by evolving the contour in the direction of steepest

descent of the energy functional and regarding as a solution a location where the contour no longer changes with time. This is accomplished mathematically by deriving a partial differential equation (PDE) from the Euler-Lagrange equations corresponding to the energy functional and then solving numerically this PDE.

This idea was first introduced in the purely boundary-based method of Snakes in [12], which were implemented using splines to parametrize the evolving curves; the fact that the number of objects detected could not dynamically change limited their applicability since a good initialization is almost always essential.

Geodesic Active Contours: One of the primary developments in the field of active contour models in the last decade has been the reformulation in [6] of the Snakes model as the computation of geodesics, with resqpect to a metric defined using the image. This gave rise to the Geodesic Active Contour model (GAC), which is parametrization free, and where the geodesic curve computation is reduced to the numerical solution of a geometric flow. This geometrical flow can be efficiently solved numerically using level set methods [26], where the contour is represented as the zero level-set of a 3D function. Thereby the evolving contours naturally split and merge, allowing the simultaneous detection of several objects. Still, a specific initialization step is necessary, where the initial curve should lie completely exterior or interior to the object boundaries.

Active Regions: Along a different line, based primarily on the (region based) Mumford-Shah functional [20], variational methods have been designed to deal with multiple cues [32] and have been later rephrased [22] in terms of geometric curve evolution schemes implemented using level set methods [26]. Specifically, in [22] the boundary and region based terms have been brought together, so that the curve evolution is driven by both statistical terms as in [32], and geometric terms as in [6]. In [7] a special case of the Mumford-Shah functional is used to give rise to purely region-based curve evolution equations which are implemented using an efficient variation of level set methods.

Unsupervised Methods, Prior Information: In [23] an unsupervised textured image segmentation method is proposed, which extends and simplifies the methods of [22]. This scheme simultaneously updates the regions and the parameters of the distributions, thereby dynamically learning the model of each region. Another active field deals with the incorporation of prior knowledge about the desired segmentation, like shape knowledge [9, 24].

Multiple Cues: Multiple cues are dealt with using a statistical model to express the feature similarity term within each region; intensity, texture, motion, etc. features are merged in a vector of random variables which is assumed to follow a region-specific multivariate distribution. The only modification of the evolution equations is then in the region-based term, which uses a high dimensional vector to determine the corresponding forces, as in [5, 22, 23, 32].

4.2 Statistical Methods

Another approach to segmentation considers the segmentation labels as a random field, and recasts the segmentation problem as a problem of deriving estimates from this field, like Maximum-A-Posteriori (MAP), or Minimum Mean Squared Error (MMSE) estimates.

Markov/Gibbs Random Fields: By assuming that each segmentation label conditioned on its neighbors is independent of the rest of the field leads to the Markov Random Field (MRF) model. These dependencies between the labels x_i, x_j at neighboring nodes in an MRF are encoded in the *clique potentials* $\Psi(x_i, x_j)$, while the influence of the observed data is encoded in the *observation potentials* $\Phi_i(x_i)$. The energy of a configuration $X = \{x_1, \ldots, x_N\}$ is then given by

$$E(X) = \prod_i \Phi_i(x_i) \prod_{x_j \in \mathcal{N}_i} \Psi(x_i, x_j) \tag{2}$$

The Gibbsian probability of X is related to its energy by Boltzmann's law:

$$P(X) = \frac{1}{Z} e^{-E(X)/T} \tag{3}$$

Efficient Inference Algorithms for MRFs: Despite their flexibility and their robustness compared to variational algorithms, MRF-based segmentation algorithms had fallen into dismay during the previous decade, since the stochastic relaxation-based algorithms that dominated research in the 1980's [11] were not fast enough for practical applications. During the recent years interest in MRFs has resurged due to the introduction of efficient algorithms for inference, based on deterministic, local computations [31]. The Belief Propagation (BP) algorithm has been used since the 1980's for inference on random fields, where the dependencies between the random variables can be expressed in terms of a graph without loops. The results obtained in this case are exact and variations of the same algorithm can be used for deriving MAP estimates and estimating the marginal distributions at nodes. Applying the BP algorithm on graphs with loops (like MRFs for low-level vision) gives rise to the Loopy Belief Propagation (LBP) algorithm which, even though not guaranteed to give the exact estimates, gives very good results in practice. Theoretical justifications for the application of the LBP algorithm have established links with the Bethe approximation from statistical physics. In [10] specially designed LBP algorithms for computer vision are proposed that are shown to result in substantial speedups.

In [2] a sampling method from Statistical Physics, Swendsen Wang Cuts, has been modified to be applicable to MRFs for computer vision. Samples are drawn from the posterior probability of the MRF's configurations efficiently, by flipping clusters of labels which significantly speeds up inference. A closely related algorithm devised for inference on MRFs is the Graph-Cuts algorithm which is described in the following section.

Generative Models: Another active field of research in statistical methods for image segmentation is the field of generative models-based segmentation (and vision in general). The idea is to construct statistical models of the appearance of each region, and then assign the observations in the image to the region that explains them best. Most important contributions in this field include [28,29,32], which have given rise to impressive segmentation results. In this work, image segmentation is posed as parameter estimation, where the parameters include a) the boundaries of the regions, b) the type of each region (e.g. texture, intensity, face regions) c) the parameters of the generative model describing the features within each region and d) the number of the regions. The observed data is the image itself, and image segmentation and parsing is interpreted as sampling from the posterior on these parameters. In case the the only regularity enforced on the segmentation labels is that of minimum boundary length, the inference of the region labels given all the other parameters can be speeded up using a variational algorithm like active regions which is then interpreted as a greedy (non-stochastic) search in the posterior distribution of labels.

Multiple Cues: The conventional statistical approach, which has also been adopted by variational methods, groups the features from all cues in a vector of random variables and has been described in the previous section. In [29] an alternative method is used to allow each region to be either a texture, a color region or any other categorical type of regions, by stochastically performing *jump moves* in the posterior distribution of the segmentation parameters. This makes it possible to choose among the cues that are used to perform segmentation in an elegant and theoretically sound way.

4.3 Graph-based Algorithms

Graph Theoretic Algorithms represent the image as a weighted graph, where the nodes are image pixels and the edge weights encode the information available about the desired segmentation: this can be either in the form of pairwise weights, like in the case of MRFs, or bottom-up knowledge, like the output of some edge-detector, which indicates that two nodes (pixels) should belong to different segments. The segmentation is determined, using techniques from graph theory, by finding a partition (cut) of the graph such that a criterion is (min-) maximized.

Graph Cuts: Graph Cuts [4] have been devised to perform efficient inference of the MAP estimate on MRFs, and they are guaranteed to converge to a solution that is at least equal to a fixed fraction of the global optimum. By a combination of swap moves the segmentation labels are changed in clusters which results in important speedups in the convergence rates. Specifically, even though the inference of the MAP estimate is NP hard, based on classical computer science algorithms for graphs, the complexity of estimating a suboptimal solution is reduced to polynomial in the product of the number labels and pixels. This

method has been also applied to minimize the cost function of the geodesic active contour model [3].

Normalized Cuts: Based on a different approach, in the Normalized Cuts algorithm [27] a graph is constructed where all the pixels-nodes in the image are connected, with weights on the graph determined by the affinity (in feature space and in location) of the pixels. The problem of segmentation is then phrased as determining a partition of the graph, where the affinity of the nodes 'lost' by separating the graph (the cut) is minimal *when divided by* the total affinities of each cluster to the whole graph (the normalization). Thereby, even though separating a single node into a cluster may result in a low cut value, when normalized this will be close to the maximum normalized cut value. The partitions are calculated by considering initially the labels of the nodes continuous, solving an eigenvector problem and then discretizing the eigenvectors. This algorithm gives state of the art results and has been extended to deal with motion segmentation and object-based segmentation.

Minimal Path Finding Methods: In [17] the image is considered as a graph where the nodes are either the pixels of the image or regions (tiles) of the mosaic (fine segmentation) of the image and the segmentation is equivalent to finding the shortest distance on the graph or the minimum spanning tree, where the distance is defined as the length of some path. Depending on the way that the path is defined, e.g shortest path, cheapest path, easiest path, different segmentation schemes occur. In [8] the problem of segmentation is investigated using a minimum path approach: the global minimum attainable by an active contour is detected between two end points, by modifying the model energy to include an internal regularization term in addition to the external potential term. In a similar manner, the proposed approach in [1] relies on some kind of energy partition of the image domain, where the energy is defined by measuring a pseudometric-based distance to a source point; thus, the choice of an energy and a set of sources determines the tessellation of the domain.

Multiple Cues: Multiple cues are integrated by expressing the weights between nodes as the product (or maximum) of the weights corresponding to each cue. For example in the normalized cuts method, affinity terms based on texture features are multiplied with affinity terms based on pixel proximity as well as edge-based affinities etc.

4.4 Morphological Methods

Mathematical morphology is a nonlinear image analysis methodology that is primarily based on set- and lattice- theoretic approaches whose goals are to quantify the geometrical structure of images. Among the morphological methodologies that have been developed for various image analysis tasks, the most prominent segmentation methodology proposed is the *watershed transform*.

The idea of the watershed-based morphological segmentation can be summa-

rized in three stages. At a first stage the image is simplified and processed in such a way that the presence of noise is reduced and useless (redundant) information is removed, thus producing an image that consists mostly of flat and large regions. The second stage involves the region-feature extraction, where the goal is to extract some special features such as small homogeneous regions, called markers, which will be used as the starting points for the flooding process. The selection of the markers is probably the most difficult task and the strategies for finding them are diverse and problem dependent; many case studies are listed in [16]. The last stage is the application of the watershed transform, where a gradient image is constructed and its topographic surface is flooded by sources placed at the position of the markers. The watershed construction grows the markers until the exact contours of the objects are found. Catchment basins without sources are flooded by already flooded neighboring catchment basins. In order to avoid the merging of lakes produced by different sources, dams are erected to keep them separated. The set of dams constitutes the boundaries of the resulting segmentation.

This model of segmentation has been extendedly studied and used during the past years. Compared to the other segmentation methods, the watershed has several advantages, including the proper handling of gaps and the placement of boundaries at the most significant edges. Despite its success it has reached its limits as new segmentation needs have appeared, which have triggered the development of a rich and coherent framework able to deal with a large variety of segmentation tasks. Today, mathematical morphology proposes a series of tools, to be used in a sequence or combined with the classical watershed model, in order to satisfy the present segmentation needs:

Multiscale segmentation: The use of multiple scales appears to be an essential part for all types of morphological segmentation whether they are automatic or interactive. A multiscale segmentation can be produced by the progressive merging of regions, and extraneous knowledge can be incorporated in this process [17]. Useful partitions can be extracted from multiscale representations of the image, with a degree of coarseness that can de determined through interaction with the user.

Generalized floodings: In [14, 15] F. Meyer presented several particular modes of flooding, in order to properly segment different types of images. Synchronous flooding is investigated, where all lakes share some common criterion, such as altitude, depth, area or volume. He also established a continuum between multiscale segmentation and segmentation with markers by using fuzzy markers: sources are placed, whose flood is slowed down by a factor associated to each marker. In [19], in order to introduce geometrical regularization constraints in a morphological segmentation scheme and make it comparable to energy-driven methods, the computation of the watershed on a smooth relief by implementing viscous flooding has been proposed.

Levelings: Without prior filtering an image contains numerous tiny and

meaningless homogeneous zones. For this reason, filtering before segmenting is often mandatory. The levelings [18] are novel object-oriented morphological filters, which enable the separation of homogeneous zones from the transition zones. Homogeneous zones can thus be enlarged, without blurring or displacing the contours: for this reason its is possible to directly segment the filtered image, with no need to go back to the original image.

PDE formulation of watershed transform: Maragos and Butt [13] explore the common theoretical concepts, tools, and numerical algorithms used in differential morphology and curve evolution. They focus on morphological operator representations for curve evolution as well as evolution laws for various morphological curve operations and distance transforms and consider them as the major route to connect differential morphology and curve evolution to the eikonal PDE. A minimum distance algorithm is derived for the watershed; in a continuous formulation, this is modelled via the eikonal PDE, which can be solved using curve evolution algorithms. They introduce the implementation of the watershed transform via the eikonal PDE. Another approach to the PDE formulation of the watershed transform was proposed in [21], where watershed segmentation is represented as an energy minimization problem, using the distance-based definition of the watershed line. A priori considerations about smoothness are then be imposed by adding the contour length to the energy function. This leads to a segmentation method called *Watersnakes*, that integrates the strengths of watershed segmentation and energy based segmentation, in a PDE formulation.

Lattice-theoretic segmentation and Connectivity: In [25] Serra investigated the problem of segmentation from the viewpoint of the algebraic lattice of image partitions. His approach is based on a new concept of connectivity defined on algebraic complete sup-generated lattices.

Multiple Cues: Multiple Cues come into play when estimating the gradient image, by using edge detectors that combine multiple cues. As an alternative, in [30] a hierarchical method is proposed where a scale-space edge detection scheme is derived by the vector-valued diffusion of multiple cues.

4.5 Benchmarks and comparatives

Only recently has a common benchmark been made available for public use: in the Berkeley Segmentation Dataset (4.6) thousands of manually segmented images have been made available, so as to systematize the comparison among image segmentation methods.

4.6 Useful URLs

Variational Methods:

The Odyssee team (PDEs for segmentation):

http://www-sop.inria.fr/odyssee/research/1/index.en.html
Level-set methods for computer vision:
http://www.math.ucla.edu/~imagers/

Statistical Methods:

The UCLA team's work on statistical methods for segmentation:
http://civs.stat.ucla.edu/Segmentation/Segment.htm
The Pattern Theory Group:
http://www.dam.brown.edu/ptg

Graph based Methods:

The Berkeley vision group segmentation project & normalized cuts:
http://www.cs.berkeley.edu/projects/vision/Grouping/overview.html
http://www.cis.upenn.edu/~jshi/
Graph cuts and belief propagation for vision:
http://www.cs.cornell.edu/vision/

Benchmarks:

The Berkeley segmentation benchmark:
http://www.cs.berkeley.edu/projects/vision/grouping/segbench/

4.7 Prior-based Segmentation in the Variational Framework

4.7.1 From Occam's Razor to Mumford-Shah

Image segmentation is a key element in the extraction of semantic content from images. It transforms the voluminous and intricate raw pixel data to a compact set of potentially meaningful segments. Classical methods for object segmentation and boundary determination rely on intrinsic local image features such as gray level values, texture features or image gradients. However, when the image to segment is noisy or taken under less than ideal illumination conditions, the use of a-priori knowledge is essential.

A-priori knowledge regarding image structure can be obtained from various sources. We distinguish between generic prior knowledge, that applies to most images, to application-specific prior knowledge. Notable examples for generic prior knowledge are the principle of parsimony (Occam's Razor) [H] and the Gestalt principles of perception [K63]. In contrast, shape priors are generally application specific.

The incorporation of a-priori knowledge in image segmentation is not trivial. The challenge is to represent the a-priori knowledge in mathematical terms that can be linked to a segmentation algorithm. For example, the generic principle of parsimony is the basis of the Minimum Description Length (MDL) [BRY98] approach. However, the seemingly appealing straightforward representation of

parsimony using Kolmogorov complexity [LV97] can lead to computationally intractable algorithms.

In their seminal paper, Mumford and Shah [MS89] casted the segmentation problem as a functional minimization problem. The functional consists of a *fidelity term*, that quantifies the discrepancy between the segmentation output and the input image data, and two additional terms that represent, in fact, the principle of parsimony. One measures the non-uniformity within segments, the other evaluates the total length of boundaries segment boundaries.

4.7.2 From Mumford-Shah to Chan-Vese

Formally, Mumford and Shah [MS89] proposed to segment an input image $f : \Omega \rightarrow \mathbb{R}$ by minimizing the functional

$$E(u, C) = \frac{1}{2}\int_{\Omega}(f-u)^2 dxdy + \lambda\frac{1}{2}\int_{\Omega - C}|\nabla u|^2 dxdy + \nu|C| \ , \tag{4}$$

simultaneously with respect to the segmenting boundary C and the piecewise smooth approximation u, of the input image f. The functional minimization problem is approached via the calculus of variations. Straightforward minimization is not trivial, due to the dependence of the integration domains on the unknown segmentation. This technical difficulty can be alleviated using, e.g., the Γ-convergence framework [AT90].

When the weight λ of the smoothness term tends to infinity, u becomes a piecewise constant approximation, $u = \{u_i\}$, of f. The functional can now be expressed as

$$E(u, C) = \frac{1}{2}\sum_i \int_{\Omega_i}(f-u_i)^2 dxdy + \nu|C| \qquad \cup_i \Omega_i = \Omega, \quad \Omega_i \cap \Omega_j = \emptyset \tag{5}$$

In the two phase case, Chan and Vese [CV01] used a level-set function $\phi \in \mathbb{R}^3$ to embed the contour $C = \{x \in \Omega |\ \phi(x) = 0\}$ [OS88], and introduced the Heaviside function $H(\phi)$ into the energy functional:

$$E_{CV}(\phi, u_+, u_-) = \int_{\Omega}\left[(f-u_+)2H(\phi) + (f-u_-)2(1-H(\phi)) + \nu|\nabla H(\phi)|\right] dxdy \tag{6}$$

where

$$H(\phi) = \begin{cases} 1 & \phi \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Using Euler-Lagrange equations for the functional (6), the following gradient descent equation for the evolution of ϕ is obtained:

$$\frac{\partial \phi}{\partial t} = \delta(\phi)\left[\nu \ \text{div}\ (\frac{\nabla\phi}{|\nabla\phi|}) - (f-u_+)2 + (f-u_-)2\right] \ . \tag{8}$$

A smooth approximation of $H(\phi)$ (and $\delta(\phi)$) must be used in practice [CV01]. The scalars u_+ and u_- are updated in alternation with the level set evolution to take the mean value of the input image f in the regions $\phi \geq 0$ and $\phi < 0$, respectively:

$$u_+ = \frac{\int f(x,y)H(\phi)dxdy}{\int H(\phi)dxdy} \qquad u_- = \frac{\int f(x,y)(1-H(\phi))dxdy}{\int (1-H(\phi))dxdy} \tag{9}$$

The bi-level Chan-Vese functional (6), provides the framework for modern prior-based segmentation techniques, that use application-specific shape priors in addition to the generic principle of parsimony. The case of polygonal contours was considered in [UYK03].

4.7.3 From generic priors to shape priors

In the presence of occlusion, shadows and low image contrast, generic prior knowledge is insufficient by itself. Prior knowledge on the shape of interest is then necessary [U96]. The recovered object boundary should then be compatible with the expected contour, in addition to being constrained by length, smoothness and fidelity to the observed image. To achieve this, the energetic formulation (6) can be extended by adding a prior shape term [CSS03]:

$$E(\phi, u_+, u_-) = E_{CV}(\phi, u_+, u_-) + \mu E_{shape}(\phi), \qquad \mu \geq 0. \tag{10}$$

The main difficulty in the integration of prior information into the variational segmentation process is the need to account for possible pose transformations between the known contour of the given object instance and the boundary in the image to be segmented. Many algorithms [CTTHW01, CKS02, CKS01, LGF00, TYWT01, LFGW00, RP02] use a comprehensive training set to account for small deformations. These methods employ various statistical approaches to characterize the probability distribution of the shapes. They then measure the similarity between the evolving object boundary (or level set function) and representatives of the training data. The performance of these methods depends on the size and coverage of the training set.

4.7.4 Selected recent contributions of MUSCLE partners

The incorporation of prior knowledge in image segmentation using the variational framework is a hot research topic. Here a examples of recent contributions of MUSCLE partners in this field.

- **INRIA-ARIANA:** M. Rochery, I. H. Jermyn and J. Zerubia have recently developed high order active contours, and applied them to the detection of line networks in satellite imagery [RJZ03].

- **TAU-VISUAL:** T. Riklin-Raviv, N. Kiryati and N. Sochen brought together concepts from variational segmentation and vision geometry. Their method is deterministic, and accounts for significant projective transformations between a *single* prior image and the image to be segmented [RKS04].

References

[1] P. A. Arbelaez and L. Cohen. Energy partitions and image segmentation. *Journal of Mathematical Imaging and Vision*, 20:4357, 2004.

[2] A. Barbu and S.C. Zhu. Graph partition by Swendsen-Wang cuts. In *Proc. International Conference on Computer Vision*, 2003.

[3] Y. Boykov and V. Kolmogorov. Computing geodesics and minimal surfaces via graph cuts. In *Proc. International Conference on Computer Vision*, 2003.

[4] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

[5] T. Brox, M. Roussonl, R. Deriche, and J. Weickert. Unsupervised segmentation incorporating colour, texture, and motion. Technical Report RR-4760, INRIA, 2003. `ftp://ftp.inria.fr/INRIA/publication/publi-pdf/RR/RR-4760.pdf`.

[6] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *International Journal of Computer Vision (IJCV)*, 22(1):61–79, 1997.

[7] T. Chan and L. Vese. Active contours without edges. *IEEE Tans. on Image Processing*, 10(2), 2001.

[8] L. D. Cohen and R. Kimmel. Global minimum for active contour models: A minimal path approach. *International Journal of Computer Vision*, 24(1):5778, 1997.

[9] D. Cremers, N. Sochen, and C. Schnorr. Multiphase dynamic labeling for variational recognition-driven image segmentation. In *Proc. European Conference on Computer Vision*, 2004.

[10] P. Felzenszwalb and D. Huttenlocher. Efficient belief propagation for early vision. In *Proc. International Conference on Computer Vision & Pattern. Recognition*, 2004.

[11] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

[12] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. In *Proc. International Conference on Computer Vision*, 1987.

[13] P. Maragos and M. A. Butt. Curve evolution, differential morphology, and distance transforms applied to multiscale and eikonal problems. *Fundamenta Informaticae*, 41:91 – 129, 2000.

[14] F. Meyer. Morphological multiscale and interactive segmentation. In *Proc. IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing*, 1999.

[15] F. Meyer. Flooding and segmentation. In *Proc. International Symposium on Mathematical Morphology*, 2000.

[16] F. Meyer and S. Beucher. Morphological segmentation. *Journal of Visual Communication and Image Representation*, 1(1):21 – 45, 1990.

[17] F. Meyer and P. Maragos. Multiscale morphological segmentations based on watershed, flooding, and eikonal PDE. In *Proc. of Scale-Space*, volume 1682 of *LNCS*, pages 351–362, 1999.

[18] F. Meyer and P. Maragos. Nonlinear scale-space representation with morphological levelings. *Journal of Visual Communication and Image Representation*, 11:245–265, 2000.

[19] F. Meyer and C. Vachier. Image segmentation based on viscuous flooding. In *Proc. International Symposium on Mathematical Morphology*, 2002.

[20] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational-problems. *Communications on Pure and Applied Mathematics*, 42(5):577–685, 1989.

[21] H. T. Nguyen, M. Worring, and R. Boomgaard. Watersnakes: energy-driven watershed segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(3):330–342, 2003.

[22] N. Paragios and R. Deriche. Geodesic active contours and level set methods for supervised texture segmentation. *Internation Journal of Computer Vision*, 2002.

[23] M. Rousson, T. Brox, and R. Deriche. Active unsupervised texture segmentation on a diffusion based space. In *Proc. International Conference on Computer Vision & Pattern Recognition*, 2003.

[24] M. Rousson and N. Paragios. Shape priors for level set representations. In *Proc. European Conference in Computer Vision*, 2002.

[25] J. Serra. Connections for sets and functions. *Fundamenta Informaticae*, 41, 2000.

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

[27] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2001.

[28] Z. Tu, X. Chen, A. Yuille, and S.C. Zhu. Image parsing: Unifying segmentation, detection, and recognition. In *Proc. International Conference on Computer Vision*, 2003.

[29] Z. Tu and S.C. Zhu. Image segmentation by data-driven MCMC. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(5):657–673, 2002.

[30] I. Vanhamel, I. Pratikakis, and H. Sahli. Multiscale gradient watersheds of color images. *IEEE Trans. on Image Processing*, 12(6):617–626, 2003.

[31] J. Yedidia, W. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. In *Proc. Uncertainty in Artificial Intelligence*, 2001. `http://www.merl.com/reports/docs/TR2001-22.pdf`.

[32] S.C. Zhu and A. Yuille. Region competition: Unifying snakes, region growing and Bayes/MDL for multiband image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(9):884–900, 1996.

References

[AT90] L. Ambrosio and V.M. Tortorelli, "Approximation of Functionals Depending on Jumps by Elliptic Functionals via Γ-Convergence", *Communications on Pure and Applied Mathematics*, Vol. XLIII, pp. 999-1036, 1990.

[AK02] G. Aubert and P. Kornprobst, *Mathematical Problems in Image Processing*, Springer, New York, 2002.

[C95] A. Chambolle, "Image Segmentation by Variational Methods: Mumford and Shah functional, and the Discrete Approximation", *SIAM Journal of Applied Mathematics*, Vol. 55, pp. 827-863, 1995.

[K63] K. Koffka, "Principles of Gestalt Psychology", Harcourt Brace, New York, 1963.

[MS89] D. Mumford and J. Shah, "Optimal Approximations by Piecewise Smooth Functions and Associated Variational Problems", *Communications on Pure and Applied Mathematics*, Vol. 42, pp. 577-684, 1989.

[H] F. Heylighen, "Occam's Razor", in: F. Heylighen, C. Joslyn and V. Turchin (editors): Principia Cybernetica Web (Principia Cybernetica, Brussels), URL: http://pespmc1.vub.ac.be/OCCAMRAZ.html.

[BRY98] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling", *IEEE Trans. Information Theory*, Vol. 44, pp. 2743-2760, 1998.

[LV97] M.Li and P. Vitanyi, "An Introduction to Kolmogorov Complexity and Its Applications", 2nd Edition, Springer Verlag, 1997.

[CV01] T.F. Chan and L.A. Vese, "Active Contours Without Edges", *IEEE Trans. Image Processing*, Vol. 10, pp. 266-277, 2001.

[OS88] S. Osher and J.A. Sethian, "Fronts Propagating with Curvature-Dependent Speed: Algorithms Based on Hamilton-Jacobi Formulations", *Journal of Computational Physics*, Vol. 79, pp. 12-49, 1988.

[U96] S. Ullman, "High-Level Vision: Object Recognition and Visual Cognition", Chapter 8, MIT Press, 1996.

[CSS03] D. Cremers, N. Sochen and D. Schnorr, "Towards Recognition-based Variational Segmentation Using Shape Priors and Dynamic Labeling", Proc. Intl. Conf. on Scale-Space Theories in Computer Vision, pp. 388-400, 2003.

[CTTHW01] Y. Chen, S. Thiruvenkadam, H.D. Tagare, F. Huang and D. Wilson, "On the Incorporation of shape priors into geometric active contours", Proc. VLSM01, pp. 145-152, 2001.

[CKS02] D. Cremers, T. Kohlberger and C. Schnorr, "Nonlinear Shape Statistics in Mumford-Shah Based Segmentation", Proc. ECCV02, Vol II, pp. 93-108, 2002.

[CKS01] D. Cremers, T. Kohlberger and C. Schnorr, "Nonlinear Shape Statistics via Kernel Spaces", Proc. DAGM01, pp. 269-276, 2001.

[LGF00] M.E. Leventon, W.E.L. Grimson and O. Faugeras, "Statistical Shape Influence in Geodesic Active Contours", Proc. CVPR00, Vol. I, pp. 316-323, 2000.

[TYWT01] A. Tsai, A. Yezzi, W.M. Wells, C. Tempany, D. Tucker, A. Fan, W.E.L. Grimson and A.S. Willsky, "Model-Based Curve Evolution Technique for Image Segmentation", Proc. CVPR01, Vol. I, pp. 463-468, 2001.

[LFGW00] M. Leventon, O. Faugeras, W. Grimson and W. Wells, "Level set based segmentation with intensity and curvature priors", Proc. Workshop on Mathematical Methods in Biomedical Image Analysis, pp. 4-11, 2000.

[RP02] M. Rousson and N. Paragios, "Shape Priors for Level Set Representations", Proc. ECCV 2002, LNCS 2351, pp. 78-92, 2002.

[RJZ03] M. Rochery, I. H. Jermyn, and J. Zerubia, "Higher Order Active Contours and their Application to the Detection of Line Networks in Satellite Imagery", Proc. IEEE Workshop Variational, Geometric and Level Set Methods in Computer Vision, at ICCV, Nice, France, October 2003.

[UYK03] G. Unal, A. Yezzi and H. Krim, "Information-Theoretic Active Polygons for Unsupervised Texture Segmentation", preprint, 2003.

[RKS04] T. Riklin-Raviv, N. Kiryati and N. Sochen, "Unlevel-Sets: Geometry and Prior-based Segmentation", Proc. 8th European Conference on Computer Vision (ECCV'2004), Prague, Czech Republic, May 2004, Part IV: *Lecture Notes in Coputer Science* #3024, pp. 50-61, Springer, 2004.

5 Saliency & Visual Feature Organization

Visual systems that have evolved in nature appear to exercise a mechanism that places emphasis upon areas in a scene without necessarily recognising objects that lie in those areas. Organisms having the benefit of vision are thereby able to sense danger and direct attention rapidly towards the unusual without having to tolerate the initial delay of a recall from memory. Treisman and Gelade [1] in their feature-integration theory make the distinction between scenes that require relatively slow focused attention to analyse and those which can be processed more rapidly during a pre-attentive stage. Evidence shows that it is relatively easy to spot a target "O" that pops out amongst a background of "N"s and "T"s, but time consuming to locate one's offspring in a school photograph. They posed, as others have done since, the question why features that distinguish a target from the background in pre-attentive vision when applied separately often do not when they appear in conjunction. Wolfe [2] emphasises that there is no clear distinction between slow serial and fast parallel mechanisms in visual search and that the evidence shows a continuum of search results in which both mechanisms perhaps play a part.

Desimone and Duncan [3] in their review confirm that strengthening the perceived grouping between targets and background objects makes the background harder to ignore. Furthermore they suggest that there is little evidence that there are separate representations for different features such as orientation and colour in the cortex stating that cells that respond to a single type of stimulus have yet to be found. They conclude that pre-attentive vision is an emergent property of competitive interactions acting in parallel across the visual field and not the binding together of a set of separate feature measures.

Nothdurft [4] has shown that the salience of targets in human vision is nearly always increased if multiple contrasts in orientation, luminance and motion are present. The addition was mostly nonlinear, which indicated that the underlying mechanisms were not independent and not separable as Desimone and Duncan suggest.

Experiments by Reinagel and Zador [5] using eye trackers show that subjects are attracted by image regions possessing high contrast and also by neighbourhoods in which pixel correlations drop off rapidly with distance. They observe that this strategy increases the entropy of the effective visual input and is in accord with measures of informativeness and cognitive surprise.

Early computational models by Koch et al [6] of attention generate maps that encode the visual environment for different elementary features such as orientation of edges and colour contrast and combine these into an overall saliency map. The most conspicuous neighbourhoods are taken to be those that give rise to the most activity in the saliency map as a result of activity in corresponding feature maps. Many authors have put emphasis upon identifying specific features that are normally associated with saliency and combining these to produce such maps.

Milanese et al [7] used five feature maps in their analysis of static scenes. After applying filters and passing the maps through a nonlinear relaxation process, they are averaged and thresholded to produce the saliency map. Itti et al [8] have defined a system which models visual search in primates. 42 features based upon linear filters and centre surround structures encoding intensity, orientation and colour, are used to construct a saliency map that reflects areas of high attention. Supervised learning is suggested as a strategy to bias the relative weights of the features in order to tune the system towards specific target detection tasks. They observed that salient objects appearing strongly in only a few dimensions may be masked by noise present in a larger number of dimensions. Han et al [25, 29] use the Itti model to determine the best positions to seed a region growing algorithm for object extraction.

Osberger and Maeder [9] identified perceptually important regions by first segmenting images into homogeneous regions and then scoring each area using a number of intuitively selected measures. The approach was limited by the success of the segmentation techniques used. Luo and Singhal [10] also devised a set of intuitive saliency features and weights and used them to segment images to depict regions of interest. The integration of the features was not attempted. Marichal et al [11] used fuzzy logic to segment object boundaries before assigning levels of interest based upon a number of criteria. Zhao et al [12] employed features reflecting size, distance from the centre of the image, boundary length, compactness and colour to determine region importance.

Reisfeld et al [27] detect symmetries in grey level images as a means of identifying certain image locations that are worthy of attention. The work relies heavily upon edge extraction and was extended to colour images by Heidemann [28].

Walker et al [13] suggested that object features that best expose saliency are those which have a low probability of being mis-classified with any other feature. Mudge et al [14] also considered the saliency of a configuration of object components to be inversely related to the frequency that those components occur elsewhere. Oliva et al [26] again use the idea that frequent features in images are more likely to belong to the background. They use orientation, scale and texture features to calculate saliency likelihoods and also incorporated contextual information.

Several models of visual attention [8, 15, 33] have their counterparts in surround suppression in primate V1 [16]. Grigorescu et al. [15] use this model and confirm qualitative explanations of visual pop-out effects. They obtain their results using pre-selected orientation sensitive Gabor energy filters, and apply their ideas to contour detection. Whereas Grigorescu et al obtain their results using pre-selected orientation sensitive Gabor energy filters, Stentiford [17, 19] takes an approach that is not so restricted and generates features appropriate to the image in question. In this way features that determine levels of attention, which may or may not be orientation dependent, are not excluded from consideration.

By observing that visual receptive fields are sensitive to orientations, scales and intensity variations, some wavelet-based salient point detectors have also been proposed [36, 37] leading to a description of the image content by a limited number of single points that are considered as perceptually relevant.

5.1 Applications

There is no doubt that the selection of salient image regions is a key problem in many fields of image processing. These cover object recognition [20], content-based image retrieval [17, 18, 21], image compression [19, 22, 31], natural image classification [38], and medical image registration [23]. Specific examples include:

- Fan et al [24] use attentional heuristics to identify sub-pictures likely to be of interest and more easily displayed while browsing on the small screen of a PDA.
- Models of visual attention have been used to extend the JPEG 2000 compression standard with some success [30].
- A blurring strategy based on visually salient regions of video frames is used to compress video signals without substantially interfering with human fixations [32].
- Measures of attention have been shown to enhance the performance of object recognisers by focusing analysis on the regions likely to contain relevant objects [31].

5.2 Discussion

Almost all researchers in the field agree that saliency plays a major part in the recognition processes that take place in the human visual system, but the exact relationship is unclear. There is also agreement that local salient features certainly include those that occur relatively rarely in a scene. Indeed it is hard to see how a feature that was present throughout an image could be visually attentive.

Many attention models make use of plausible features such as orientation, intensity, colour and texture as a first step in the identification of regions of interest. Concepts of centre surround suppression are often applied that emphasise the importance of local differences and lead to an attention map that displays the relative saliency of various regions in the image.

However, there is no evidence that visual systems make use of a small number of predefined features that we might intuitively believe to be important in attention mechanisms. The favoured few certainly characterise saliency in many images, but there is an infinitude of other possible feature combinations to choose

from and there will always therefore be images whose saliency is not captured. Salient features are most likely to be different in different images and the diverse possibilities emerge only at the time the images are viewed. This means that implementations of attention mechanisms that use any form of pre-defined feature measurements may preclude solutions in the search space and be unable to handle unseen material.

The same problem arises again in recognition where we are seeking features in common between two or more patterns. There is no guarantee that a pre-selected feature set and associated representations will encompass the inter-pattern similarities necessary to obtain a satisfactory performance, although sensible feature design based on prior knowledge will always yield good results. Indeed it is a criticism that techniques in feature extraction and pattern classification are treated separately with the result that features are not constructed according to classification performance [34].

One of the most difficult problems in saliency-based image representation is to design efficient local descriptors that aim at describing the image content in the neighborhood of each salient area. Usually, these descriptors are designed by hand and are inspired from intuitive considerations. However, there is no evidence that the resulting descriptor provides a sufficiently good image representation and/or a good discriminative power for an image recognition task. Finally, when an image is represented by a set of salient zones, the natural order between pixel values is lost and the image can no longer be seen as a vector, increasing thus the complexity of the matching step for assessing the similarity between images or objects of interest. Classically, image registration approaches are used but in these cases, salient zones are considered independently of each others. However, human eyes are able to recognize a scene from a set of focus of attention and saccadic eye movements, showing that the information is accumulated during these movements. A registration approach is not able to reach this property since the local information contained in each salient feature ignores the local information localized in other places of the image. To simulate the image categorization from saccadic movements, a solution has been proposed in [38] that consists in linking the detected salient features to obtain an attributed string of local descriptors. An order between salient features is thus recovered and the matching step can be performed thanks to a string-edit distance.

References

[1]A. M. Treisman and G. Gelade, "A feature-integration theory of attention", Cognitive Psychology, 12, pp. 97-136, 1980.

[2]J. M. Wolfe, "Visual Search in Attention" edited by H Pashler, Psy-

chology Press, 1998.

[3]R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention", Annu. Rev. Neurosci., 18, pp. 193-222, 1995.

[4]H. C. Nothdurft, "Salience from feature contrast: additivity across dimensions", Vision Research, 40, pp. 1183-1201, 2000.

[5]P. Reinagel and M. Zador, "Natural scene statistics at the centre of gaze", Network: Comp. Neural Syst., 10, pp. 341-350, 1999.

[6]C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry", Human Neurobiol., 4, pp. 219-227, 1985.

[7]R. Milanese, S. Gil and T. Pun, "Attentive mechanisms for dynamic and static scene analysis", Optical Engineering, 34, No 8, pp. 2420-2434, August 1995.

[8]L. Itti, C. Koch and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis", IEEE Trans PAMI, 20, No 11, pp. 1254-1259, November 1998.

[9]W. Osberger and A. J. Maeder, "Automatic identification of perceptually important regions in an image", 14th IEEE Int. Conference on Pattern Recognition, 16-20th August 1998.

[10]J. Luo and A. Singhal, "On measuring low-level saliency in photographic images", IEEE Conf. On Computer Vision and Pattern Recognition, June 2000.

[11]X. Marichal, T. Delmot, C. De Vleeschouwer, V. Warscotte and B. Macq, "Automatic detection of interest areas of an image or of a sequence of images", IEEE Int. Conf on Image processing, 1996.

[12]J. Zhao, Y. Shimazu, K. Ohta, R. Hayasaka, and Y. Matsushita, "An outstandingness oriented image segmentation and its application", Int. Symposium on Signal Processing and its Applications, August 1996.

[13]K. N. Walker, T. F. Cootes and C. J. Taylor, "Locating Salient Object Features", British Machine Vision Conference, 1998.

[14]T. N. Mudge, J. L. Turney and R. A. Volz, "Automatic generation

of salient features for the recognition of partially occluded parts", Robotica, 5, pp. 117-127, 1987.

[15]C. Grigorescu, N. Petkov and M.A. Westenberg, "Contour detection based on nonclassical receptive field inhibition," IEEE Trans on Image Processing, vil12, no 7, pp 729-739, 2003.

[16] H-C. Nothdurft, J. L. Gallant and D. C. Van Essen, "Response modulation by texture surround in primate area V1: Correlates of "popout" under anesthesia," Visual Neuroscience, 16, pp 15-34, 1999.

[17]Stentiford F W M: 'An attention based similarity measure with application to content based information retrieval', SPIE Vol 5021, Storage and Retrieval for Media Databases, Santa Clara (22-24, January 2003).

[18]Pauwels E J and Frederix G: 'Finding Salient Regions in images: Non-Parametric clustering for image segmentation and grouping', Computer Vision and Image Understanding, 75(1, 2):73-85 (August 1998).

[19]F. W. M. Stentiford, "An estimator for visual attention through competitive novelty with application to image compression", 22nd Picture Coding Symposium, Seoul, April 2001.

[20]R.M. Haralick and L.G. Shapiro, Computer and Robot Vision, vol 2, 1993.

[21]Q. Tian, N. Sebe, M.S. Lew, E. Loupias, and T.S. Huang, "Image retrieval using wavelet-based salient points," J. Electronic Imaging, vol 10, no 4, pp. 835-849, 2001.

[22]C.M. Privitera and L.W. Stark, "Algorithms for defining visual regions of interest: comparison with eye fixations," IEEE Trans Pattern Analysis and Machine Intelligence, vol 22, no 9, pp.970-982, 2000.

[23]J. Maintz and M. Viergever, "A survey of medical image registration," Medical Image Analysis, vol 2, no 1, pp. 1-36, 1998.

[24]X. Fan, X. Xie, W.Y. Ma, H.J. Zhang and H.Q. Zhou, "Visual attention based image browsing on mobile devices," ICME 2003.

[25]J. Han, M. Li, H Zhang and L. Guo, "Automatic attention object extraction from images," ICIP 2003.

[26]A. Oliva, A. Torralba, M.S. Castelhano and J.M. Henderson, "Top-down control of visual attention in object detection," ICIP 2003.

[27]D. Reisfeld, H. Wolfson and Y. Yeshurun, "Context-free attentional operators: the generalised symmetry transform," Int. J. Computer Vision, vol 14, pp. 119-130, 1995.

[28]G. Heidemann, "Focus-of-attention from local color symmetries," IEEE Trans Pattern Analysis and Machine Intelligence, vol26, no 7, July 2004.

[29]J. Han, K.N. Ngan, M. Li and H. Zhang, "Towards unsupervised attention object extraction by integrating visual attention and object growing," ICIP 2004.

[30]A. Nguyen, V. Chandran, S. Sridharan and R. Prandolini, "Guidelines to using region of interest coding in JPEG 2000," Proc. Int. Symposium on Digital Signal Processing and Communication Systems (DSPCS), Gold Coast, Australia, 8-11 December, 2003.

[31]Y. Hu, X Xie, Z. Chen and W.Y. Ma, "Attention model based progressive image transmission," ICME 2004.

[32]L. Itti, "Automatic attention-based prioritization of unconstrained video for compression," In: Proc. SPIE Human Vision and Electronic Imaging IX (HVEI04), San Jose, CA, (B. Rogowitz, T. N. Pappas Ed.), Vol. 5292, pp. 272-283, Jan 2004.

[33]O. Le Meur, D. Thoreau, E. Francois, D. Barba and P. Le Collet, " From low-level perception to high-perception level: a coherent approach for VA modelling," In: Proc. SPIE Human Vision and Electronic Imaging IX (HVEI04), San Jose, CA, (B. Rogowitz, T. N. Pappas Ed.), Vol. 5292, pp. 284-295, Jan 2004.

[34]Duda R.O., Hart P.E. and Stork D.G., "Pattern Classification," John Wiley, NY, 2001.

[35]F. W. M. Stentiford, N. Morley and A. Curnow, "Automatic identification of regions of interest with application to the quantification of DNA damage in cells," Proc SPIE Vol 4662, San Jose, 20-26 Jan 2002.

[36]E. Loupias, N. Sebe, S. Bres and J.M. Jolion, "Wavelet-based Salient Points for Image Retrieval", In: Proc of the IEEE International Conference

on Image Processing, Vancouver, September 2000.

[37]C. Laurent, N. Laurent, M. Visani, "Color Image Retrieval Based on Wavelet Salient Features Detection", In: Third International Workshop on Content-Based Multimedia Indexing, Rennes, pp. 327-334, September 2003.

[38]J. Ros, C. Laurent, "Natural Image Classification Using Foveal Strings", In: The International Workshop on Image, Video, and Audio Retrieval and Mining, Sherbrooke, October 2004.

6 Image Sequence Features

Motion is seamlessly perceived by human beings when directly observing a day-life scene, but also when watching films, videos or TV programs, or even various domain-specific image sequences such as meteorological or heart ultrasound ones. However, motion information is hidden in the image sequences supplied by image sensors. It has to be recovered from the observations formed by the image intensities in the successive frames of the sequence.

Assumptions (i.e., *data models*) must be formulated to relate the observed image intensities with motion. When dealing with video, the commonly used data model is the brightness constancy constraint which states that the intensity does not change along the trajectory of the moving point in the image plane (at least, to a short time extent). The motion constraint equation can then be expressed in a differential form that relates the 2D velocity vector, the spatial image gradient and the temporal intensity derivative at any point p in the image. Nevertheless, this enables to locally retrieve one component of the velocity vector only, the so-called normal flow, which corresponds to the aperture problem. Then, other constraints (i.e., *motion models*) must be added. They are supposed to formalize known, expected or learned properties of the motion field, and this implies to somehow introduce spatial coherence or more generally contextual information.

Visual motion information can involve different kinds of mathematical variables. First, we can deal with *continuous variables* to represent the motion field : velocity vectors $\mathbf{w}(p)$ with $\mathbf{w}(p) \in \mathbb{R}^2$, or parametric motion models with parameters $\theta \in \mathbb{R}^d$ with d denoting the number of parameters. Let us note that the latter can be equivalently represented by the model flow vectors $\{\mathbf{w}_\theta(p)\}$ with $\mathbf{w}_\theta(p) \in \mathbb{R}^2$. Second, we can consider *discrete values or symbolic labels* to code motion detection output: binary values $\{0, 1\}$, or motion segmentation output: number n of the motion region or layer with $n \in \{1, ..., N\}$.

Spatial coherence can be formalized by conditional densities defined on local neighborhoods as in Markov Random Field (MRF) models, or equivalently by potentials on cliques as in Gibbs distributions. Another way is to first segment each image into spatial regions according to a given criterion (grey level, colour, texture) and to analyse the motion information over these regions. Perceptual grouping schemes can also be envisaged.

6.1 Local motion measurements

The brightness constancy assumption along the trajectory of a moving point $p(t)$ in the image plane, with $p(t) = (x(t), y(t))$, can be expressed as $dI(x(t), y(t), t)/dt = 0$, with I denoting the image intensity function. By applying the chain rule, we get the well-known motion constraint equation [23, 36]:

$$r(p,t) = \mathbf{w}(p,t).\nabla I(p,t) + I_t(p,t) = 0 \ , \tag{1}$$

where ∇I denotes the spatial gradient of the intensity, with $\nabla I = (I_x, I_y)$, and I_t its partial temporal derivative. The above equation can be straightforwardly extended to the case where a parametric motion model is considered, and we can write:

$$r_\theta(p,t) = \mathbf{w}_\theta(p,t).\nabla I(p,t) + I_t(p,t) = 0 \ , \tag{2}$$

where θ denotes the vector of motion model parameters. It can be easily derived from equation (1) that the motion information which can be locally recovered at a pixel p is contained in the *normal flow* given by:

$$\nu(p,t) = \frac{-I_t(p,t)}{\|\nabla I(p,t)\|} \ . \tag{3}$$

It can also be written in a vectorial form: $\boldsymbol{\nu}(p,t) = \frac{-I_t(p,t)}{\|\nabla I(p,t)\|}\boldsymbol{\omega}_{\nabla I}(p,t)$, where $\boldsymbol{\omega}_{\nabla I}$ denotes the unit vector parallel to the intensity spatial gradient. However, it should be clear that the orientation of the normal flow vector does not convey any information on the motion direction, but implicitly on the object texture (for inner points) or on the object shape (for points on the object border). Besides, the normal flow can be computed at the right scale to enforce reliability as explained in [17].

In case of a moving camera and assuming that the dominant image motion is due to the camera motion and can be correctly described by a 2D parametric motion model, we can exhibit the *residual normal flow* given by:

$$\nu_{res}(p,t) = \frac{-DFD_{\hat{\theta}}(p,t)}{\|\nabla I(p,t)\|} \ , \tag{4}$$

where $DFD_{\hat{\theta}}(p,t) = I(p + \mathbf{w}_{\hat{\theta}}, t+1) - I(p,t)$ is the displaced frame difference corresponding to the compensation of the dominant motion described by the estimated motion model parameters $\hat{\theta}$.

Since the computation of intensity derivatives is usually affected by noise and can be unreliable in nearly uniform areas, it may be preferable to consider the local mean of the absolute magnitude of normal residual flows weighted by the square of the norm of the spatial intensity gradient (as proposed in [25, 42]):

$$\bar{\nu}_{res}(p,t) = \frac{\sum_{q\in\mathcal{F}(p)} \|\nabla I(q,t)\|.|DFD_{\hat{\theta}_t}(q)|}{\max\left(\eta^2, \sum_{q\in\mathcal{F}(p)} \|\nabla I(q,t)\|^2\right)} \ , \tag{5}$$

where $\mathcal{F}(p)$ is a local spatial window centered in pixel p (typically a 3×3 window), and η^2 is a predetermined constant related to the noise level. An interesting property of the local motion quantity $\bar{\nu}_{res}(p)$ is that the reliability of the conveyed motion information can be locally evaluated. Given the lowest motion magnitude δ to be detected, we can derive two bounds, $l_\delta(p)$ and $L_\delta(p)$, verifying the following properties [42]. If $\bar{\nu}_{res}(p) < l_\delta(p)$, the magnitude of the (unknown) true velocity

vector $\mathbf{w}(p)$ is necessarily lower than δ. Conversely, if $\bar{\nu}_{res}(p) > L_\delta(p)$, $\|\mathbf{w}(p)\|$ is necessarily greater than δ. The two bounds l_δ and L_δ can be directly computed from the spatial derivatives of the intensity function within the window $\mathcal{F}(p)$.

By defining the motion quantity $\bar{\nu}_{res}(p)$, we already advocate the interest of considering spatial coherence to compute motion information. Here, it simply amounts to a weighted averaging over a small spatial support and it only concerns the data model. In the same vein, more information can be locally extracted by considering small spatio-temporal supports, either through spatio-temporal (frequency-based) velocity-tuned filters as in [18] or using 3D orientation tensors [3,37]. On the other hand, more benefit can be gained by introducing contextual information through the motion models.

6.2 Motion detection and segmentation

Previous approaches to motion detection can be split into two categories: methods based on motion segmentation and methods thresholding image differences. A synopsis of methods from both categories is proposed in [35].

Motion segmentation methods require an accurate estimation of the 2D apparent motion in the image. This is not trivial since computing motion estimation on the various image supports arising from objects in the scene is definitely demanding. However, some specific problems cannot be solved without information on the orientation of motion. For example, apparent motion estimation enables to conclude that the motion of a static background and a static foreground, although different in their 2D projection, are induced by the same camera motion and should therefore not be detected as independently moving. This can be done using parallax and rigidity constraints [26], [38], [54], [53] and [12].

Thresholding methods are applied to temporal image differences. The temporal image differences of static regions and those of moving objects can be statistically modeled. Within the class of thresholding methods, different kinds of image structures can be considered. They may range from decision on single pixels to spatial regularization using MRFs or active contours. The highest level of structure is representing moving objects by their spatio-temporal envelopes. A different form of structure is to consider only edges and to detect those that are moving.

A review of basic methods is given in [28]. Likelihood tests for motion detection were early introduced in [24]. More complex methods are proposed in [47] for modeling the spatial distribution of either noise or signal and selecting an appropriate threshold. Markov Random Fields are a natural extension in order to introduce contextual information into the detection scheme [1]. Previous to image differencing, camera-induced motion can also be estimated and compensated. To this end, a 2D parametric motion model is used in [42] along with hierarchical MRFs. In [13] wavelet analysis and robust techniques are introduced to estimate dominant motion and hierarchical MRFs are again exploited.

A higher level of spatial structure can be reached using active contour and the level set theory. For instance the approach described in [44] applies active contours for detection and tracking of moving objects. It relies on the gray level intensities for providing object boundaries. Two methods using level sets implementation are introduced in [31]: one purely based on motion and another enforcing correspondence between motion boundaries and intensity boundaries. Besides, they can distinguish between different moving objects.

Bayesian approaches like MRF and variational methods both rely on energy minimization techniques. In the Bayesian framework, the aim is to maximize the posterior probability of a model given the observations. Variational techniques minimize an energy functional yielding a contour evolving according to some constraints. The formulation of energies for both approaches is similar. It consists of a regularization term and a term to fit observations. However, it is not possible to assess the validity of the extremum. In other words, it is not possible to interpret the value of the extremum. Thus, the best state is reached given the model and the observations, but the quality of this state cannot be assessed.

Temporal integration improves quality and stability of the detection. Accumulating motion information over time makes moving objects more salient with regard to noise. This enables to detect small slowly moving objects. Directionally consistent flow is accumulated over time in [59]. A graph to represent moving objects is exploited in [7], and object trajectories are optimal paths in this graph. Spatio-temporal image intensity gradients to create mosaics of the background are used in [46]. Residual motion is then propagated and accumulated without optical flow computation. A threshold allows one to balance between false alarms and minimal detectable motion. It is not clear how to set this threshold, unless empirically.

As mentioned above, motion information accumulates on edges. The work in [51] concentrates on these highly contrasted features to detect multiple layered motions. In [27] contour fragments are matched. The different transformations issued from those matchings are clustered into background and moving objects. This work is closer to shape matching issues than to motion detection. Both methods rely on a Canny-type edge detector.

Some work has also been devoted to distinguish between motion and changes. Changes are variations of image intensities which do not correspond to a real moving object : shadows and reflections but also aliasing [6].

The use of perceptual criteria for change detection appears in [30] and [49]. In [30] an *a contrario* framework is applied to detect changes in satellite images of urban landscapes. The considered local change information is the image gradient orientation. In [49], the use of perceptual organization is considered to build the spatio-temporal envelopes of moving objects. The approach is essentially applied to human undergoing fronto-parallel motion in front of a static camera. The envelopes localize the motion information but do not provide shape information. In [55], camera-induced motion is first compensated using 2D parametric motion

models. The perceptual grouping principle allows the computation of automatic detection thresholds. Detection operates on three frames only. Boundaries of moving objects are retrieved through an image segmentation based on meaningful intensity level lines. Moreover, a confidence level for each detected region is derived through the so-called number of false alarms evaluated according to the *a contrario* model. The lower this number, the more reliable the detected event.

Let us now consider motion segmentation meant as the competitive partitioning of the image into motion-based homogeneous regions. One important step ahead in solving the motion segmentation problem was to formulate the motion segmentation problem as a statistical contextual labeling problem or in other words as a discrete Bayesian inference problem. Segmenting the moving objects is then equivalent to assigning the proper (symbolic) label (i.e., the region number) to each pixel in the image, while estimating 2D parametric (usually affine) motion models over the region supports (which is obviously a chicken-and-egg problem). The advantages of this class of methods are mainly two-fold. Determining the support of each region is then implicit and easy to handle: it merely results from extracting the connected components of pixels with the same label. Introducing spatial coherence can be straightforwardly (and locally) expressed by exploiting MRF models. This formulation can also encompass the determination of motion layers by assuming that the regions of same label are not necessarily connected [50].

Specifying (simple) MRF models at a pixel level (i.e., sites are pixels and a 4- or 8-neighbour system is considered) is efficient, but remains limited to express more sophisticated properties on region geometry (e.g., more global shape information [10]) or to handle extended spatial interaction. Multigrid MRF models [22] (as used in [42, 43]) is a means to address somewhat the second concern (and also to speed up the minimization process while usually supplying better results). An alternative is to first segment the image into spatial regions (based on grey level, colour or texture) and to specify a MRF model on the resulting graph of adjacent regions as done in [19]. The motion region labels are then assigned to the nodes of the graph (which are the sites considered in that case). This enables to exploit more elaborated and less local *a priori* information on the geometry of the regions and their motion. However, the spatial segmentation stage is often time consuming, and getting an effective improvement on the final motion segmentation accuracy remains questionable. Using the level-set framework is another way to precisely locate region boundaries while dealing with topology changes [44], but handling a competitive motion partioning of the image (with the number of regions *a priori* unknown) remains an open issue in that context even if recent attempts have been reported [11, 31].

Let us also mention other recent work on Bayesian motion segmentation, exploring the use of edge motion [51], offering extension to spatio-temporal models

[11], or introducing (two-step) hidden Markov measure field (HMMF) models [32]. Tensor voting could also be considered as an implicit way to enforce spatial coherence [40].

6.3 Optical flow computation

By definition, the velocity field formed by continuous vector variables is a complete representation of the motion information. Computing optical flow based on the data model of equation (1) requires to add a motion model enforcing the expected spatial properties of the motion field, that is, to resort to a regularization method. Such properties of spatial coherence (more specifically, piecewise continuity of the motion field) can be expressed on local spatial neighborhoods. First methods to estimate discontinuous optical flow fields were based on MRF models associated with Bayesian inference [21,35,52] (i.e., minimization of a discretized energy function). Then, continuous-domain models were designed based on PDE formalism [2, 8, 29, 56]. Spatial coherence can also be explicitly formulated by first segmenting the image in spatial regions forming the delimited domains where motion models, either dense or parametric ones, can be defined and estimated [5, 19].

A general formulation of the global (discretized) energy function to be minimized to estimate the velocity field $\mathbf{w}$ can be given by:

$$E(\mathbf{w},\zeta) = \sum_{p\in S} \rho_1[r(p)] \; + \sum_{p\sim q} \rho_2[\|\mathbf{w}(p)-\mathbf{w}(q)\|, \zeta(p'_{p\sim q})] + \sum_{A\in\chi} \rho_3(\zeta_A) \;\; , \qquad (6)$$

where S designates the set of pixel sites, $r(p)$ is defined in (1), $S' = \{p'\}$ the set of discontinuity sites located midway between the pixel sites and χ is the set of cliques associated with the neighborhood system chosen on S'. In [21], quadratic functions were used and the motion discontinuities were handled by introducing a binary line process ζ. Then, robust estimators were popularized [4, 33] leading to the introduction of so-called auxiliary variables ζ now taking their values in $[0, 1]$. Depending on the followed approach, the third term of the energy $E(\mathbf{w}, \zeta)$ can be optional. Multigrid MRF are moreover involved in the scheme developed in [33]. Besides, multiresolution incremental schemes are required to compute optical flow in case of large displacements. Dense optical flow and parametric motion models can also be jointly considered and estimated, which enables to supply a segmented velocity field as designed in [34].

Recent advances have dealt with the computation of fluid motion fields involving the definition of a new data model (derived from the continuity equation of the fluid mechanics) and of a motion model preserving the underlying physics of the visualized fluid flows (2^{nd} order div-curl constraint) as defined in [9]. A comprehensive investigation of physics-based data models is described in [20].

6.4 Motion recognition

Exploiting the tremendous amount of multimedia data, and specifically video data, requires to develop methods able to extract information at a higher level (more semantic) level. Video summarization, video retrieval or video surveillance are examples of applications. Inferring concepts from low-level video features is a highly challenging problem. The characteristics of a semantic event have to be expressed in terms of video primitives (color, texture, motion, shape ...) sufficiently discriminant w.r.t. content. This remains an open problem at the source of active research activities.

In [57], statistical models for components of the video structure are introduced to classify video sequences into different genres. The analysis of image motion is widely exploited for the segmentation of videos into meaningful units or for event recognition. Efficient motion characterization can be derived from the optical flow, as in [48] for human action change detection. In [60], the authors use very simple local spatio-temporal measurements, i.e., histograms of the spatial and temporal intensity gradients, to cluster temporal dynamic events. In [58], a principal component representation of activity parameters (such as translation, rotation ...) learnt from a set of examples is introduced. The considered application was the recognition of particular human motions, assuming an initial segmentation of the body.

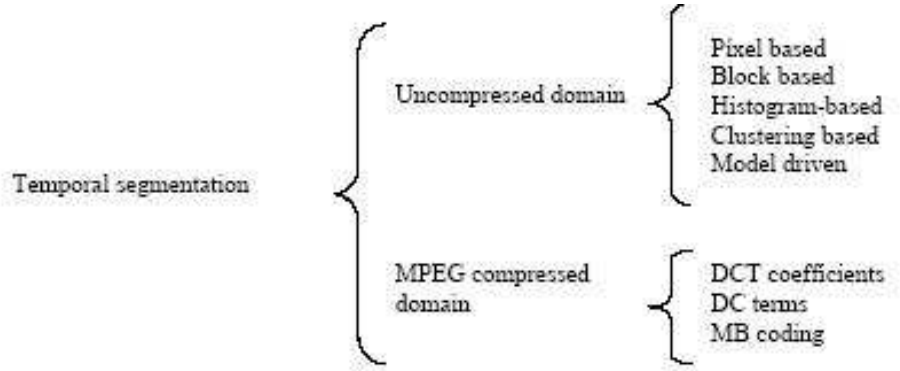
In [14], video abstraction relies on a measure of fidelity of a set of key-frames and a measure of summarizability derived from MPEG-7 descriptors. In [39], spatio-temporal slices extracted in the volume formed by the image sequence are exploited for clustering and retrieving tasks. Sport videos are receiving specific attention due to the economical importance of sport TV programs and to future services to be designed in that context. Different approaches have been recently investigated to detect highlights in sports videos [15].

In [45], a statistical approach is proposed involving modeling, (supervised) learning and classification issues, to deal with concepts related to events in videos, more precisely, to dynamic content. Focus is put on motion information. Since no analytical motion models are available to account for the variety of dynamic contents to be found in videos, motion models have to be specified and learned from the image data. To this end, new mixed-state probabilistic motion models are introduced. Furthermore, such a probabilistic modelling allows the derivation of a parsimonious motion representation while coping with errors in the motion measurements and with variability in a given kind of motion content. Scene motion (i.e., the residual image motion) and camera motion (i.e., the dominant image motion) are handled in a distinct way. Indeed, these two sources of motion bring important, different but complementary, information which have to be explicitly taken into account for event detection. As for motion measurements, on one hand, 2D parametric motion models are considered to capture the camera motion, and on the other hand, low-level local motion features to account for the

scene motion. They convey more information than those used in [60], while still easily computable contrary to optic flow.

6.5 Video shot detection algorithms

During the last years several papers describing the state of the art in video shot detection have been published, like [BR96], [DAE95] or [KC01]. We would like to provide an extended review of current techniques describing the most relevant approaches in view of our applications in the context of video analysis for information retrieval. A well-accepted classification of shot boundary detection algorithms is based on the format of the input sequence: uncompressed (raw) images, or an MPEG bit-stream. Under these main categories, we group the different techniques in the state of the art, as represented in the figure below.



6.5.1 Segmentation in the compressed domain

Most of existing algorithms work in the uncompressed domain. They use as input data a sequence of frames, being the color of the pixels the only information directly available. Commonly, transitions are detected by means of a threshold over a similarity measure. According to the metric used to compute the distance between frames and the technique used in the segmentation process, current algorithms can be classified into five sub-groups [KC01].

Pixel comparison :
This type of transition detection algorithms, also known as template matching, relies on the computation of pixel-wise difference between consecutive frames. One of the earliest methods was implemented by Kikukawa and Kawafuchi in [KK92]. They compute the difference between images as the absolute sum of pixel differences. For gray-level images, the frame difference curve FD(t) is the absolute sum of pixel differences between frames at t and t + 1. By extension, for color images, the frame difference curve is computed as the mean of the FD

in each color component. The proposed algorithm detects a transition when the value of the frame difference exceeds a certain threshold. The main drawback of this approach is that the system is not able to distinguish between a small change of the whole image and a large change in a small portion of the image (due to object motion, for example) resulting in a high rate of false positives. An improvement for this technique is presented by Zhang et al. [ZKS93]. In [HJW94], Hampapur et al. compute what they call chromatic images by dividing the change in gray level of corresponding pixels from two consecutive images by the value of that pixel in the last image. Then, fades should correspond to uniform chromatic images. But this approach is also very sensitive to camera and object motion.

Campisi et al. [CNS02] present a classic but robust method. A difference image sequence is computed from the video. Then each difference image is segmented into blocks. The similarity between consecutive collocated blocks is computed and the sum of the similarities for the whole image, and within a temporal window, is compared with a dynamic threshold to detect a fade or a dissolve.

An approach based on the variance of pixel intensities was proposed by Alattar [Ala97]. Fades were detected first by recording all negative spikes in the time series of the second order difference of the pixel intensity variance, and then by ensuring that the first order difference of the mean of the video sequence was relatively constant next to the negative spike. A combination of both approaches is described in Truong et al. [TDV00]. These methods have a relatively high false detection rate.

One algorithm for cuts and one for fades and dissolves have been developed by Huang and Liao [HL01]. The first one checks if the value of the DC image difference has a local maximum (within a window) at the current frame transition. Then the ratio of this maximum to the next highest maximum in the window is compared with two thresholds. If it is higher than the highest threshold, a cut is detected. If it is between the two thresholds, a heuristic algorithm that considers the histogram difference and the static or dynamic nature of the appearing and disappearing shots (based on edge differences) is used. In the algorithm for gradual transitions, the image difference from a specific frame is differentiated, its zero-crossings are computed and then low-pass filtered. Then the transition is the interval where there are few resultant zero-crossings, meaning that the difference is monotonically increasing.

Lienhart [Lie99] proposed first to locate all monochromatic frames in the video as potential start/end points of fades. Monochromatic frames were identified as frames with standard deviation of pixel intensities close to zero. Fades were then detected by starting to search in both directions for a linear increase in the standard deviation of pixel intensity/color.An average hit rate of 87% was reported at false alarm rate of 30%.

Block-based pixel comparison :
The previous techniques are based on global image characteristics, which makes them very sensitive to camera and object motion. In order to increase the robustness, block-based approaches use local characteristics of the image. Under this approach the frame difference is computed comparing the B blocks in which the image is divided. Following this strategy, different techniques can be derived depending on the computation of the difference between blocks. Kasturi and Jain [KJ91] compute the similarity between blocks using a likelihood ratio based on the mean and the variance of the luminance values in each block. Blocks are then counted as changed if and only if the likelihood ratio exceeds a certain threshold. When the percentage of changed blocks is above a second threshold, a cut is declared. This method is more robust in front of slow camera and object motion, but it still presents a high ratio of false positives. Moreover, because statistical values must be computed, its computational burden is higher than for template matching approaches. In order to improve the robustness in front of motion, Shahraray [S95] proposes a method where several matches are considered for each block. A new algorithm called net comparison is presented by Xiong et al. in [XLI95]. As previous techniques, it declares a cut according to the percentage of changed blocks between consecutive frames, but in order to increase the processing speed only a part of the image is analyzed. As an extension, in [XL98] the same authors propose to sub-sample in both time and space. The last technique we present based on block differences was developed by Demarty [9]. In the first step, her approach computes the difference between consecutive frames (based on pixel value differences) and applies a block-based criterion to determine the amount of change in each block, creating what the author calls the transition mask. Then, a global criterion is applied on the transition mask in order to determine the amount of change between consecutive frames. By applying the same procedure all over the sequence, a monodimensional curve is created. This curve is filtered using morphological operators (a modified version of the top hat) so that transitions can be detected by imposing a threshold. While this approach has proven to be robust for the detection of scene cuts, it is less capable to detect gradual transitions because of the thresholding of the last step. Moreover, according to the author, it is sensible to abrupt changes on the images, causing that flashes or rapid motion yield false alarms.

Histogram comparison :
The main idea behind the use of the histogram is to further reduce the sensibility to the camera motion. Since histograms represent a distribution of colors, they are robust in front of rotations or changes in the viewing angle [S93]. Although in theory it is possible that completely different images present identical histograms, this is a very unlikely situation in real scenarios. As we have seen in the case of pixel comparisons, histogram-based approaches can also operate over the whole image (global histogram) or by regions (local histograms).

Global histogram comparison :
First techniques based on histogram comparisons [NT92], [T91], [ZKS93], [YYWL95] used the same approaches we have previously described for pixel-based algorithms. However, instead of using a distance based on pixel values, they define a new metric based on histogram comparisons. The histogram difference between two gray-level images at t and t + 1 can be formulated as:

$$HD(t) = \frac{1}{M} \sum m = 0^{M-1} \|h_{t+1}(m) - h_t(m)\| \quad (7)$$

where M is the total number of bins of the histogram. On that basis, a cut is declared when the value of the histogram difference $HD(t)$ exceeds a certain threshold. In spite of its simplicity, such approaches provided promising results. In order to improve the performance of these algorithms, Ueda et al. [UMY91] and Zhang et al. [ZKS93] take into account color information by means of color histograms. An important factor to take into account when histograms are compared is the color space used to represent colors. Both Smith [S97] and Gargi et al. [GOKDK95] analyze among others, the following color spaces: RGB, HSV, YUV. Up to now, the histogram-based methods are only suited to detect abrupt transitions (cuts) by using a single threshold. The algorithm proposed in [ZKS93] by Zhang et al. implements a technique called twin comparison. This algorithm also relies on the difference between color histograms but it uses two thresholds: one to detect cuts, and another to detect gradual transitions. Tests conducted by Borezcky [BR96] show that this approach significantly reduces the number of false positives. However, it slightly reduces the number of detected transitions, too.

Local histogram comparison :
The use of local histograms (computed over a portion of the image) intends to recover part of the spatial information lost from the pixel domain. The idea is to take profit of the robustness of the histograms in front of movement while introducing some spatial information to improve the performance. A block-based approach is presented by Nagasaka and Tanaka in [NT92]. Images are split into 16 non-overlapping blocks. A color histogram is computed for each block and compared to that of the corresponding block. The largest differences are discarded to be more tolerant to object and camera motion. Then, as we have seen in the case of pixel-based block comparisons, a two-threshold approach is used to detect the transitions. The first threshold is used to decide if a block is counted as changed between consecutive frames. The second threshold specifies the minimum number of changed blocks to declare a transition. This technique reduces the number of missed transitions compared to global histogram approaches at the expense of a higher rate of false alarms. Another technique based on local histograms, named selective HSV histograms, is presented by Lee and Ip in [LI94].

Clustering-based temporal video segmentation :
Thus far, the algorithms we have reviewed rely on the thresholding of a simi-

larity measure between consecutive frames. In the following, we discuss some techniques based on different approaches. In [GFT98], Gnsel et al. present an unsupervised clustering technique based on the K-Means algorithm. According to this technique, the segmentation is viewed as a 2-class clustering problem, where each transition between consecutive frames must be classified as shot change or no shot change using the K-Means algorithm. Observe that although frame transitions are labelled individually, gradual shot changes can be naturally detected: when several successive frame transitions are marked as shot change, the whole set corresponds to a gradual change. The information used to measure similarities is based on color histograms both in the YUV and RGB color spaces. Observe that this technique does not classify the different types of gradual transitions. However it overcomes one of the main drawbacks of previous techniques: the dependence of the threshold on the type of sequence. The main advantage of the clustering-based segmentation is that it allows multiple features to be simultaneously used. In [FT98], histogram and pixel-based features are jointly used.
The method of Qi et al. [QHL03] uses two binary classifiers trained with manually labelled data: one for separating cut form non-cut frames and one for separating abrupt from gradual cuts. These are either k-nearest-neighbor classifiers, naive bayesian networks or support vector machines. The gradual/abrupt classifier input is preprocessed with wavelet smoothing. Classifier input is a vector composed of whole-frame and blockwise histogram differences from 30 neighbor frames, camera motion estimations derived from MPEG2 features and blackness of the frame. The value of the work lies in the comparison of the different types of classifiers for the task, and the fact that the experimental results are on the standard TRECVID 2001 data. The optimal F-score of 0.94 for cuts and 0.70 for gradual transitions is obtained from the k-nearest-neighbor classifiers.

A novel idea was developed by Lienhart [Lie01] where dissolves (and *only* dissolves) are detected by a learning classifier (specifically a neural network). The classifier detects possible dissolves at multiple temporal scales, and merges the results using a winner-take all strategy. The interesting part is that the classifier is trained using a *dissolve synthesizer* which creates artificial dissolves from any available set of video sequences. Although the videos used for experimental verification are non-standard, performance is shown to be superior to the simple edge-based methods commonly used for dissolve detection.

Feature-based temporal video segmentation :
Another category of techniques, which are not based on the thresholding of a frame similarity curve, bases the segmentation process on the analysis of the evolution of a certain feature. In [BBB98], [BG96], [GB98], Bouthemy et al. present several algorithms based on motion analysis (global or local). According to their reasoning, in a real video sequence the camera and object motion are continuous in the shots. However, each transition represents a discontinuity in such motion. This approach declares a transition each time there is a change in the motion model. While it is quite robust in front of movement, it is only able to

detect abrupt transitions. Relaying on the same principle of motion continuity, in [ZMM99], [ZMM95] Zabih et al. present a technique based on the analysis of contour pixels. A cut is declared when most of the edges change; a progressive change of the edges should correspond to gradual transitions.

Heng and Ngan [HN01] present an advanced edge-based approach. They extract the edges from the video and refine them. Then they match them across frames using a complex technique which includes, among other things, exhaustive matching, various constraints, motion estimating through clustering, dilation of edges, and edge joint tracking. Matched edges transfer their identity to edges of the next frame, and thus it is possible to determine the backward lifetime (age) of each edge. Each frame's majority object lifetime is the near-maximum lifetime of edges in the frame. If it is below a threshold, a gradual transition is detected. If it is near zero, a cut is detected. In addition, in order to detect the type of gradual transition the frame is segmented into blocks and the forward and backward lifetimes of the edges in each block are checked to determine if they belong to the appearing or disappearing shot.

Li et al. define the Joint Probability Image of two frames as a matrix where the element $[i, j]$ contains the probability that a pixel with color i in the first image has color j in the second image. They also derive the Joint Probability Projection Vector (JPPV) and the Joint Probability Projection Centroid (JPPC) as 1-dimensional and scalar statistics of the JPI. Specific types of transition (dissolve, fade, dither) are observed to have specific patterns of JPI. To detect shot changes, possible locations are first detected by using the JPPC, and they are then refined and validated by matching the JPPV with patterns of specific transitions.

Zhao and Grosky [ZG03] split each frame into blocks and compute the average hue and saturation for each. For every hue and saturation value, the angles of the Delaunay triangulation of the blocks that have this value is found and the angles of the triangulation are histogrammed. The feature vector for each frame is 10 hue values $\times$ 36 angle bins + 10 saturation values $\times$ 36 angle bins = 720 values. This is compared to the standard HS histogram method with 10 (H) $\times$ 10 (S) = 100 bins. Optionally, latent semantic indexing is applied. This is essentially a SVD dimensionality reduction of the frame vectors which is followed by normalization and frequency based weighting. A shot change is detected if the magnitude of the difference of the feature vectors (histogram, anglogram, or LSI of either) is above a threshold T_2. If it is between two thresholds T_1 and T_2 it is verified semi-automatically. The biggest drawback of the work, however, is the necessity of user interaction for the detection of gradual transitions.

The algorithm of Zhang et al. [ZCS03], called *PixSO*, begins by calculating the classic sum of pixel differences. If this above a threshold, a cut is detected. If it is between a high and a low threshold, each frame is segmented into two classes (foreground and background) by an unsupervised segmentation algorithm and

the change between these classes is computed. If it is above a threshold, objects are defined as connected components of foreground. The object correspondence between frames is computed based on distance of centroids. If overlap between corresponding objects is above a threshold, a cut is detected. The authors claim 0.96 recall and 0.92 precision.

Model-driven temporal video segmentation :
All the techniques we have surveyed so far extract some kind of data from the video sequence and then analyze it in order to detect some pattern which corresponds to shot transitions. These techniques are known as data-driven. Alternatively, model-driven approaches tackle the problem of video segmentation from a different point of view. They characterize the different types of transition by means of mathematical models. Hampapur et al. [HJW95] create models for different types of transition. Their method takes into account the editing process of video sequences to characterize the transitions. For instance, dissolves are modeled as the weighted average of frames from the shots before and after the transition. The segmentation algorithm tries to locate those frames of the sequence that fit the model to declare a transition. As other previously discussed techniques, it is quite sensitive to motion. In [AJ94] Aigrain and Joly introduce what they call differential model of motion picture. Instead of modeling the value of image pixels across transitions, they define a model to characterize the difference between them. The technique presented by Yu et al. [YBH97] is a mixture between data- and model-driven approaches. At the first step, an histogram-based algorithm is applied to detect cuts. Then, fades and dissolves are detected by inspection of the frames in-between. Their model is based on the evolution of the number of edge pixels across gradual transitions. A different technique is discussed in [BW98], where Boreczky and Wilcox use hidden Markov models (HMM) to detect shot transitions. They generate a three component vector for each pair of consecutive frames, based on color (gray-level histogram difference), motion (estimate of object motion) and audio information. They use the sequence of vectors to feed a six state HMM (shot, cut, fade, dissolve, pan and zoom). The main advantage of this algorithm is that thanks to the training phase (inherent to any HMM) there is no need for thresholds.

Shot detection is performed by Janvier et al. [JBMMP03] in three independent steps. First the appearance of a cut is detected by estimating the probability that a boundary exists between two frames. The various probability functions are derived experimentally and their parameters are computed from training data and the derived "shots" are further segmented into pieces that exhibit greater continuity using a linear model and dynamic programming with minimum message length criterion to fit this model. In order to give acceptable results in the task of shot detection, these segments are then merged using a 2-class K-means. This operates on color histograms and mutual information [CNP02] of frames within a window around the segment change and classifies them into "shot change" and

"no shot change" categories.

A thorough analysis of the shot detection problem is presented by Hanjalic [Han02], and a probabilistically based algorithm is described. A discontinuity value between frames is defined as the average difference between average YUV components of matching blocks within these frames. For detecting abrupt transitions adjacent frames are compared, while for gradual ones frames that are apart by the minimum shot length are compared. The a priori likelihood functions of the discontinuity values are obtained by experiments on manually labelled data. Their choice is largely heuristic. The a priori probability of shot detection conditional on time from last shot detection is modelled as a Poisson function, which was observed to offer good results. The shot detection probability is refined with a heuristic derived from the pattern of the discontinuity values (for cuts and fades) and in-frame variances (for dissolves) in a window around the current frame. Effectively a different detector is implemented for each transition type.

Sánchez and Binefa [SB03] model each shot as a Coupled Markov Chain (CMC), which encodes the probabilities of temporal change of a set of features. Here, hue and motion of a 16x16 image block are used. For each frame, two CMCs are computed, one from the shot up to and including that frame, and one from that frame and the next one. Then the Kullback-Leibler divergence between the two distributions is compared with a threshold adaptively (but heuristically) computed from the mean and standard deviation of the changes encoded in the first CMC. The fact that the dynamics of the disappearing shot are inherently taken into consideration for detecting shot changes is what gives this method its strength and elegance.

Subspace-based video segmentation :
A feature-independent method is presented by Liu and Chen [LC02]. They use a modified PCA algorithm on each shot to extract its principal eigenspace. For this, two novel algorithms are presented to facilitate the gradual computation of the eigenspace, and also to place greater weight on the most recent frames (by heuristic weights). A cut is detected when the difference between a frame and its projection into the eigenspace is above a static threshold.

In order to make the feature space more discriminant, Cernekova et al. [CKP03] propose performing SVD is on the color histograms of each frame to produce reduced dimension feature vectors. Shot change is detected by comparing the angle between the feature vector of each frame and the average of the feature vectors of the current shot. Shots whose feature vectors exhibit high sparseness are considered to depict a gradual transition between two shots. The main problem with this approach is the static threshold which is applied on the angle between vectors to detect shot change, which may be problematic in the case of large variation in intra-frame content and small variations in inter-frame content.

6.5.2 Segmentation in the MPEG compressed domain

Due to the increasing amount of material stored in MPEG format, performing temporal segmentation in the compressed domain presents several advantages. By avoiding to decode the analyzed sequence, the processing speed is increased and the storage requirements are greatly reduced. Moreover, operations should be faster because the lower gross data rate of compressed video. Another important advantage is that MPEG bit-streams contain several features that are not present in the raw sequence, such as motion vectors or average gray intensities. In the following sections, we review some of these techniques grouped by the type of information they use: DCT coefficients, DC terms or macro-block (MB) coding type.

Temporal segmentation based on DCT coefficients :
The main goal of the first approaches working on the compressed domain was speed. In [AHC93], Arman et al. introduce a cut detection algorithm based on the comparison of the DCT coefficients of corresponding blocks from consecutive I frames by means of a normalized inner product. In order to increase the processing speed, only a subset of coefficients from a subset of blocks is taken into account. The technique presented by Zhang et al. in [ZLGS94] is an adaptation of the well-known techniques that apply on the pixel domain. Here, on compressed video, the pixel-wise comparison is used to compute the difference between DCT coefficients of corresponding blocks. Both of the above techniques only take into account I frames because they are the only ones for which the DCT coefficients are directly available. On the one hand, this allows a great processing speed. On the other hand, however, it greatly reduces the performance of the algorithms. The loss of temporal resolution makes them even more sensible to object and camera motion, increasing the number of false positives compared to the techniques working on the uncompressed domain. Besides, gradual transitions cannot be detected.

Temporal segmentation based on DC terms :
A large set of techniques working on the compressed domain are based on the processing of DC terms. These values correspond to the mean luminance of each block of the coded image. The full set of DC terms constitutes a low resolution version of the original image, called DC-image. These techniques use basically the same approaches based on pixel [YL95] and histogram [PS97] differences we have seen in the previous sections, applied over the DC-images. The main problem here is how to efficiently extract these images from the input bit-stream. I frames are easy to handle because DCT coefficients are directly available. However, DC images for P and B frames are very costly to compute since motion vectors must be taken into account. In [SD95], Shen and Delp propose a fast reconstruction of color DC-images using an approximated approach. In [TD98], Taskiran and Delp present an extension of the previous technique based on a generalized sequence trace, which is the evolution of the difference between the feature vector of con-

secutive frames. In order to reduce the complexity of the DC-image creation, Sethi and Patel [SP95] present an algorithm where only I frames are processed. Ardizzone et al. [AGCM96] also propose an approach relying only on I frames.

An advanced statistical approach is presented by Lelescu and Schonfeld [LS03]. This additionally benefits from operating on MPEG-compressed videos. They extract luminance and chrominance for each block in every I and P frame, and then perform PCA on the resulting vectors. It should be noted that the eigenvectors are computed based only on the M first frames of each shot. The resulting vectors are modelled by a gaussian random distribution. Then the mean and covariance matrix of the distribution are computed, also in the M first frames of each shot. The originality of the approach is that a change statistic is computed for each new frame by maximum likelihood methodology (the Generalized Likelihood Ratio algorithm) and if it exceeds an experimentally determined threshold a new shot is started. The estimation can be either additively or non-additively based. The algorithm is tested on a number of videos originating from news, music clips, sports and other types where gradual transitions are common. The best results of 0.92 recall and 0.72 precision were recorded for the additively-based algorithm.

Temporal segmentation using MB coding mode and other features : Another feature from the compressed domain used in several segmentation algorithms is the coding mode of the macro-blocks (MB). For I images all the MB are coded in intra mode, while for P and B images they may be coded either intra- or inter-mode. For those portions of the sequence with small changes, most of the blocks are coded in inter-mode because it is possible to find similar blocks in the preceding frames. However, across a transition there is a lot of change and hence the number of MB coded in intra-mode should be much higher. Notice that this only applies for P and B frames. Different techniques make use of this information, usually combined with other features, as for instance [MJC95], [ZLS95], [KC98] or [FLM96]. The approaches described in Little et al. and Deardoff et al. [KC98] use as input data a sequence encoded using the M-JPEG format and searches for differences in the size of the encoded image to detect cuts. More precisely, a cut is detected when there is a sudden change on the size of the coded image. One of the main drawbacks of compressed domain techniques is the dependence of the performance on the input bit-stream itself. First, it implies a lack of generality of the algorithms because the information contained in the bit-stream depends on the coding method. Second, even when the same standard is used, sequences encoded with different encoders may lead to significant differences in performance [PS97]. Additionally, the performance achieved using compressed domain techniques is lower than that achieved using uncompressed domain approaches, especially for gradual transitions.

References

[AJ94] P. Aigrain and P. Joly. The automatic real-time analysis of film editing and transition effects and its applications. Computers & Graphics, 18(1):93-103, January-February 1994.

[AGCM96] E. Ardizzone, G.A.M. Gioiello, M. La Cascia, and D. Molinell. A real-time neural approach to scene cut detection. In I.K. Sethi and R.C. Jain, editors, Proceedings of IS&T/SPIE Conference on Storage and Retrieval for Image and Video Databases IV, San Jose, CA, USA, January-February 1996.

[AHC93] F. Arman, A. Hsu, and M.-Y. Chiu. Image processing on compressed data for large video databases. In Proceedings of the First ACM Internation Conference on Multimedia, pages 267-272, 1993.

[BBB98] S. Benayoun, H. Bernard, P. Bertolino, P. Bouthemy, M. Gelgon, R. Mohr, C. Schmid, and F. Spindler. Structuration de video pour des interfaces de consultation avancees. In 4mes Journes d'Etudes et d'Echanges Compression et Reprsentation des Signaux Audio-Visuels, CORESA'98, June 1998.

[BR96] J.S. Boreczky and L.A. Rowe. Comparison of video shot boundary detection techniques. In I.K. Sethi and R.C. Jain, editors, Proceedings of IS&T/SPIE Conference on Storage and Retrieval for Image and Video Databases IV, volume 2670, pages 170-179, San Jose, CA, USA, January-February 1996.

[BW98] J.S. Boreczky and L.D. Wilcox. A hidden markov model framework for video segmentation using audio and image features. In IEEE Proceedings of the International Conference on Acoustics, Speech and Signal Processing, ICASSP'98, volume 6, pages 3741-3744, Seattle, WA, USA, May 1998.

[BG96] P. Bouthemy and F. Ganansia. Video partitioning and camera motion characterization for content-based video indexing. In Proceedings of the IEEE International Conference on Image Processing, ICIP'96, volume 1, pages 905-908, Lausanne, Switzerland, September 1996.

[DAE95] A. Dailianas, R. B. Allen, and P. England. Comparison of automatic video segmentation algorithms. In SPIE Proceedings, Integration Issues in Large Commercial Media Delivery Systems, pages 2-16, Philadelphia, PA, USA, October 1995.

[D00] C.-H. Demarty. Segmentation et Structuration d'un Document Vido pour la Caractrisation et l'Indexation de son Contenu Smantique. PhD thesis, Ecole Nationale Suprieure des Mines de Paris, Paris, France, January 2000.

[FLM96] J. Feng, K.-T. Lo, and Mehrpour H. Scene change detection algorithm for MPEG video sequence. In Proceedings of the IEEE International Conference on Image Processing, ICIP'96, Lausanne, Switzerland, 1996.

[FT98] A.M. Ferman and A.M. Tekalp. Efficient filtering and clustering for temporal video segmentation and visual summarization. Journal on Visual Communications and Image Representation, 9(4):-351, 1998.

[GOKDK95] U. Gargi, S. Oswald, D. Kosiba, S. Devadiga, and R. Kasturi. Evaluation of video sequence indexing and hierarchical video indexing. In SPIE Proceedings, Storage and Retrieval for Image and Video Databases, pages 1522-1530, 1995.

[GB98] M. Gelgon and P. Bouthemy. Determining a structured spatio-temporal representation of video content for efficient visualization and indexing. In 5th European Conference on Computer Vision, ECCV'98, volume 1, pages 595-609, Freiburg, Germany, June 1998.

[GFT98] B. Gunsel, A.M. Ferman, and A.M. Tekalp. Temporal video segmentation algorithm using unsupervised clustering and semantic object tracking. Journal on Electronic Imaging, 7(3):592-604, 1998.

[HJW94] A. Hampapur, R. Jain, and T.E.Weymouth. Digital video segmentation. In Proceedings of the ACM Multimedia Conference and Exposition (verificar), pages 357-364, San Francisco, CA, USA, October 1994.

[HJW95] A. Hampapur, R. Jain, and T.E. Weymouth. Production model based digital video segmentation. Journal of Multimedia Tools and Applications, 1(1):9-46, March 1995.

[KJ91] R. Kasturi and R. Jain. Dynamic vision. In R. Kasturi and R. Jain, editors, Computer Vision: Principles, IEEE Computer Society Press, pages 469-480, Whashington DC, USA, 1991.

[KK92] T. Kikukawa and S. Kawafuchi. Development of an automatic

summary editing system for the audio-visual resources. Transactions on Electronic Information. J75-A, pages 204-212, 1992.

[KC98] I. Koprinska and S. Carrato. Detecting and classifying video shot booundaries in MPEG compressed sequences. In Proceedings of IX European Signal Processing Conference, EUSIPCO' 98, pages 1729-1732, Rodhes, Grece, 1998.

[KC01] I. Koprinska and S. Carrato. Temporal video segmentation: A survey. Signal Processing: Image communications, 16(5):477-500, January 2001.

[LI94] C.-M. Lee and D. M.-C. Ip. A robusts approach for camera break detection in color video sequences. In Proceedings of IAPR Workshop Machine Vision Applications, pages 502-505, Kawasaki, Japan, 1994.

[MJC95] J. Meng, Y. Juan, and S.-F. Chang. Scene change detection in a MPEG compressed video sequence. In Proceedings of IS&T/SPIE International Symposium on Electronic Imaging, volume 2417, pages 14-25, San Jose, CA, USA, 1995.

[NT92] A. Nagasaka and Y. Tanaka. Automatic video indexing and full-video search for object appeareances. In E. Knuth and L.M. Wegner, editors, Visual Database Systems II, pages 113-127, Amsterdam, 1992. Elsevier.

[PS97] N.V. Patel and I.K. Sethi. Video shot detection and characterization for video databases. Pattern Recognition, 30:583-592, 1997.

[SP95] I.K. Sethi and N.V. Patel. A statistical approach to scene change detection. In Proceedings of the IS&T/SPIE Conference on Storage and Retrieval for Image and Video Databases III, volume 2420, pages 2-11, San Jose, CA, USA, 1995.

[S95] B. Shahraray. Scene change detection and content-based sampling of video sequences. In A. Rodriguez, R. Safranek, and E. Delp, editors, Proceedings of IS&T/SPIE on Digital Video Compression: Algorithms and Technologies, volume 2419, pages 2-13, February 1995.

[SD95] K. Shen and E. Delp. A fast algorithm for video parsing using MPEG compressed sequences. In Proceedings of the IEEE International Conference on Image Processing, ICIP'95, pages 252-255, Washington DC, USA, October 1995.

[S97] J.R. Smith. Integrated Spatial and Feature Image Systems: Retrieval, Analysis and Compression. PhD thesis, Columbia University, New York City, NY, USA, 1997.

[S93] M.J. Swain. Interactive indexing into image databases. In SPIE Proceedings, Storage and Retrieval for Image and Video Databases, pages 173-187, 1993.

[TD98] C. Taskiran and E. Delp. Video scene change detecion using the generalized sequence trace. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98, pages 2961-2964, Seattle, WA, USA, 1998.

[T91] Y. Tonomura. Video handling based on structures information for hypermedia systems. In Proceedings of the ACM International Conference on Multimedia Information Systems, pages 333-344, 1991.

[UMY91] H. Ueda, T. Miyatake, and S. Yoshizawa. IMPACT: An alteractive natural-motion-picture dedicated multimedia authoring system. In Proceedings of the ACM Conference on Human Interfaces (veficicar), pages 343-350, New Orleans, Louisiana, USA, April-May 1991.

[XL98] W. Xiong and J.C.-M. Lee. Efficient scene change detection and camera motion annotation for video classification. Computer Vision and Image Understanding, 71(2):166-181, 1998.

[XLI95] W. Xiong, J.C.-M. Lee, and D. M.-C. Ip. Net comparison: a fast and effective method for classifying image sequences. In SPIE Proceedings, Storage and Retrieval for Image and Video Databases III, volume 2420, pages 318-328, San Jose, CA, USA, 1995.

[YL95] B. Yeo and B. Liu. Rapid scene analysis on compressed video. IEEE Transactions on Circuits and Systems and Video Technology, 5(6):533-544, 1995.

[YYWL95] M.M. Yeung, B.L. Yeo, W. Wolf, and B. Liu. Video browsing using clustering and scene transitions on compressed sequences. In Proceedings of the IS&T/SPIE Conference on Multimedia Computing and Networking, volume 2417, 1995.

[YBH97] H. Yu, G. Bozdagi, and S. Harrington. Featured-based hierarchical video segmentation. In Proceedings of the IEEE International

Conference on Image Processing, ICIP'97, pages 497-501, Santa Barbara, CA, USA, October 1997.

[ZMM99] R. Zabih, J. Miler, and K. Mai. A feature-based algorithm for detecting and classifying production effects. Multimedia Systems, 7:119-128, 1999.

[ZMM95] R. Zabih, J. Miller, and K. Mai. A feature-based algorithm for detecting and classifying scene breaks. In Proceedings of the ACM International Conference on Multimedia, San Francisco, CA, USA, November 1995.

[ZKS93] H.J. Zhang, A. KanKanhalli, and S.W. Smoliar. Automatic partitioning of full motion video. Multimedia Systems, 1(1):10-28, 1993.

[ZLGS94] H.J. Zhang, C.Y. Low, Y.H. Gong, and S.W. Smoliar. Video parsing using compressed data. In Proceedings of SPIE Conference on Image and Video Processing II, pages 142-149, 1994.

[ZLS95] H.J. Zhang, C.Y. Low, and S.W. Smoliar. Video parsing and browsing using compressed data. Multimedia Tools and Applications, 1:89-111, 1995.

References

[Ala97] A. M. Alattar. Detecting fade regions in uncompressed video sequences. In *Proc. 1997, IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pages 3025–3028. 1997.

[CKP03] Z. Cernekova, C. Kotropoulos, and I. Pitas. Video shot segmentation using singular value decomposition. In *International Conference on Multimedia and Expo*. Jul 2003.

[CNP02] Z. Cernekova, C. Nikou, and I. Pitas. Shot detection in video sequences using entropy-based metrics. In *International Conference on Image Processing*, pages 0 – 0. Oct 2002.

[CNS02] P. Campisi, A. Neri, and L. Sorgi. Automatic dissolve and fade detection for video sequences. In *International Conference on Digital Signal Processing*. Jul 2002.

[Han02] A. Hanjalic. Shot-boundary detection: Unraveled and resolved? *IEEE Transactions on Circuits and Systems for Video Technology*, 12(3), Feb 2002.

[HL01] C.-L. Huang and B.-Y. Liao. A robust scene-change detection method for video segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(13):1281 – 1288, Dec 2001.

[HN01] W. Heng and K. Ngan. An object-based shot boundary detection using edge tracing and tracking. *Journal of Visual Communication and Image Representation*, 12(4):217 – 239, Sep 2001.

[JBMMP03] B. Janvier, E. Bruno, S. Marchand-Maillet, and T. Pun. Information-theoretic framework for the joint temporal partitioning and representation of video data. In *International Workshop on Content-Based Multimedia Indexing*. Sep 2003.

[LC02] X. Liu and T. Chen. Shot boundary detection using temporal statistics modeling. In *International Conference on Acoustics, Speech and Signal Processing*. May 2002.

[Lie99] R. Lienhart. Comparison of automatic shot boundary detection algorithms. In *Proc. of SPIE Storage and Retrieval for Image and Video Databases VII, San Jose, CA, U.S.A.*, volume 3656, pages 290–301. January 1999.

[Lie01] R. Lienhart. Reliable dissolve detection. In *SPIE Storage and Retrieval for Media Databases*. Jan 2001.

[LS03] D. Lelescu and D. Schonfeld. Statistical sequential analysis for real-time video scene change detection on compressed multimedia bitstream. *IEEE Transactions on Multimedia*, 5(2), Mar 2003.

[QHL03] Y. Qi, A. Hauptmann, and T. Liu. Supervised classification for video shot segmentation. In *International Conference on Multimedia and Expo*. Jul 2003.

[SB03] J. Sánchez and X. Binefa. Shot segmentation using a coupled Markov chains representation of video contents. In *Iberian Conference on Pattern Recognition and Image Analysis*. Jun 2003.

[TDV00] B. T. Truong, C. Dorai, and S. Venkatesh. New enhancements to cut, fade, and dissolve detection processes in video segmentation. In *ACM Multimedia 2000*, pages 219–227. November 2000.

[ZCS03] C. Zhang, S.-C. Chen, and M.-L. Shyu. Pixso: A system for video shot detection. In *Pacific-Rim Conference On Multimedia*. Dec 2003.

[ZG03] R. Zhao and W. I. Grosky. Video shot detection using color anglogram and latent semantic indexing: From contents to semantics. In *Handbook of Video Databases: Design and Applications*. Jan 2003.

References

[1] T. Aach and A. Kaup. Bayesian algorithms for change detection in image sequences using Markov random fields. *Signal Processing: Image Communication*, 7(2):147–160, 1995.

[2] L. Alvarez, J. Weickert, J. Sánchez. Reliable estimation of dense optical flow fields with large displacements. International Journal of Computer Vision, 39(1):41-56, 2000.

[3] J. Bign, G.H. Granlund, J. Wiklund. Multidimensional orientation estimation with applications to texture analysis and optical flow. IEEE Trans. on Pattern Analysis and Machine Intelligence, 13(8) :775-790, August 1991.

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

[5] M.J. Black, A.D. Jepson. Estimating optical flow in segmented images using variable-order parametric models with local deformations. IEEE Trans. on Pattern Analysis and Machine Intelligence, 18(10):972-986, October 1996.

[6] M. J. Black, D. J. Fleet, and Y. Yacoob. Robustly estimating changes in image appearance. *Computer Vision and Image Understanding*, 78(1):8–31, 2000.

[7] I. Cohen and G. Medioni. Detecting and tracking moving objects for video surveillance. In *IEEE Conf. Computer Vision and Pattern Recognition*, Fort Collins CO, June 1999.

[8] I. Cohen, I. Herlin. Non uniform multiresolution method for optical flow and phase portrait models: Environmental application. IJCV, 33(1):29-49, September 1999.

[9] T. Corpetti, E. Mmin, P. Prez. Dense estimation of fluid flows. IEEE Trans. on Pattern Analysis and Machine Intelligence, 24(3):365-380, March 2002.

[10] D. Cremers, T. Kohlberger, C. Schnrr. Nonlinear shape statistics in Mumford-Shah based segmentation. 7th European Conference on Computer Vision, ECCV'2002, Copenhagen, Vol. LNCS 2351, Springer Verlag, 2002.

[11] D. Cremers, S. Soatto. Variational space-time motion segmentation. Proc. 9th IEEE Int. Conf. on Computer Vision, ICCV'2003, Nice, October 2003.

[12] G. Csurka and P. Bouthemy. Direct identification of moving objects and background from 2D motion models. In *7th International Conference on Computer Vision*, pages 566–571, Kerkyra, Greece, 1999.

[13] C. Demonceaux and D. Kachi-Akkouche. Motion detection using wavelet analysis and hierarchical Markov models. In *First International Workshop on Spatial Coherence for Visual Motion Analysis*, Prague, May 2004.

[14] A. Divakaran, R. Radhakrishnan, and K.A. Peker. Motion activity-based extraction of key-frame from video shots. *ICIP'02*, Rochester, Sept. 2002.

[15] A. Ekin, A.M. Tekalp, and R. Mehrotra. Automatic soccer video analysis and summarization. *IEEE Int. Trans. on Image Processing*, 12(7):796–807, July 2003.

[16] R. Fablet, P. Bouthemy, and P. Pérez. Non-parametric motion characterization using causal probabilistic models for video indexing and retrieval. *IEEE Trans. on Image Processing*, 11(4):393–407, 2002.

[17] R. Fablet, P. Bouthemy. Motion recognition using non parametric image motion models estimated from temporal and multiscale cooccurrence statistics. IEEE Trans. on Pattern Analysis and Machine Intelligence, 25(12):1619-1624, December 2003.

[18] D.J. Fleet, A.D. Jepson, Computation of component image velocity from local phase information. International Journal of Computer Vision, 5(1):77-104, 1990.

[19] M. Gelgon, P. Bouthemy. A region-level motion-based graph representation and labeling for tracking a spatial image partition. Pattern Recognition, 33(4):725-745, April 2000.

[20] H.W. Haussecker, D.J. Fleet. Estimating optical flow with physical models of brightness variation. IEEE Trans. on Pattern Analysis and Machine Intelligence 23(6):661-673, 2001.

[21] F. Heitz, P. Bouthemy. Multimodal estimation of discontinuous optical flow using Markov random fields. IEEE Trans. on PAMI, 15(12):1217-1232, December 1993.

[22] F. Heitz, P. Prez, P. Bouthemy. Multiscale minimization of global energy functions in some visual recovery problems. CVGIP : Image Understanding, 59(1):125-134, January 1994.

[23] B.K.P. Horn, B.G. Schunck. Determing optical flow. Art. Intelligence, 17:185-203, 1981.

[24] Y. Z. Hsu, H.-H. Nagel, and G. Rekers. New likelihood test methods for change detection in image sequences. *Computer Vision, Graphics, and Image Processing*, 26:73–106, 1984.

[25] M. Irani, B. Rousso, S. Peleg. Computing occluding and transparent motion. International Journal of Computer Vision, 12(1):5-16, 1994.

[26] M. Irani and P. Anandan. A unified approach to moving object detection in 2D and 3D scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(6):577–589, June 1999.

[27] V. Jain, B. B. Kimia, and J. L. Mundy. Segregation of moving objects using elastic matching. In *Workshop on Spatial Coherence for Visual Motion Analysis*, Prague, May 2004.

[28] J. Konrad. Motion detection and estimation. In A. C. Bovik, editor, *Handbook of Image and Video Processing*. Academic Press, 2000.

[29] P. Kornprobst, R. Deriche, G. Aubert Image sequence analysis via partial differential equations. Journal of Mathematical Imaging and Vision, 11(1):5-26, 1999.

[30] J. L. Lisani and J.-M. Morel. Detection of major changes in satellite images. In *IEEE International Conference on Image Processing*, Barcelona, September 2003.

[31] A.-R. Mansouri and J. Konrad. Multiple motion segmentation with level sets. *IEEE Transactions on Image Processing*, 12(2):201–220, 2003.

[32] J.-L. Marroquin, E.A. Santana, S. Botello. Hidden Markov measure field models for image segmentation. IEEE Trans. on PAMI, 25(11):1380-1387, November 2003.

[33] E. Mmin, P. Prez. Optical flow estimation and object-based segmentation with robust techniques. IEEE Trans. on Image Processing, 7(5):703-719, May 1998.

[34] E. Mmin, P. Prez. Hierarchical estimation and segmentation of dense motion fields. Int. Journal of Computer Vision, 46(2):129-155, February 2002.

[35] A. Mitiche and P. Bouthemy. Computation and analysis of image motion: A synopsis of current problems and methods. *International Journal of Computer Vision*, 19(1):29–55, 1996.

[36] H.-H. Nagel. On the estimation of optic flow: Relations between different approaches and some new results. Artificial Intelligence, 33:299-324, 1987.

[37] H.-H. Nagel, A. Gehrke. Spatiotemporally adaptive estimation and segmentation of OF-fields. 5th Eur. Conf. on Comp. Vis., ECCV'98, Freiburg, Vol. LNCS 1407, Springer, 1998.

[38] R. C. Nelson. Qualitative detection of motion by a moving observer. *International Journal of Computer Vision*, 7(1):33–46, 1991.

[39] C-W. Ngo, T-C. Pong, and H-J. Zhang. On clustering and retrieval of video shots through temporal slices analysis. *IEEE Trans. Multimedia*, 4(4):446–458, Dec. 2002.

[40] M. Nicolescu, G. Medioni. Layered 4D representation and voting for grouping from motion. IEEE Trans. on Pattern Analysis and Machine Intelligence, 25(4):492-501, April 2003.

[41] J-M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *J. of Visual Comm. and Image Repr.*, 6(4):348–365, Dec. 1995.

[42] J. Odobez and P. Bouthemy. Separation of moving regions from background in an image sequence acquired with a mobile camera. In H. Li, S. Sun, and H. Derin, editors, *Video Data Compression for Multimedia Computing*, chapter 8, pages 283–311. Kluwer Academic Publisher, 1997.

[43] J-M. Odobez, P. Bouthemy. Direct incremental model-based image motion segmentation for video analysis. Signal Processing, 6(2):143-155, 1998.

[44] N. Paragios and R. Deriche. Geodesic active contour and level sets for the detection and tracking of moving objects. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(3):266–280, March 2000.

[45] G. Piriou, P. Bouthemy and J-F. Yao. Extraction of semantic dynamic content from videos with probabilistic motion models. *Proc. European Conf. on Computer Vision, ECCV'04*, Prague, May 2004.

[46] R. Pless, T. Brodksy, and Y. Aloimonos. Detecting independent motion: The statistics of temporal continuity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):768–773, August 2000.

[47] P. L. Rosin. Thresholding for change detection. *Computer Vision and Image Understanding*, 86:79–95, May 2002.

[48] Y. Rui and P. Anandan. Segmenting visual actions based on spatio-temporal motion patterns. *CVPR'2000*, Hilton Head, SC, 2000.

[49] S. Sarkar, D. Majchrzak, and K. Korimilli. Perceptual organization based computational model for robust segmentation of moving objects. *Computer Vision and Image Understanding*, 86:141–170, 2002.

[50] H. S. Sawhney, S. Ayer Compact representations of videos through dominant and multiple motion estimation. IEEE Trans. on PAMI, 18(8):814-830, August 1996.

[51] P. Smith, T. Drummond, and R. Cipolla. Layered motion segmentation and depth ordering by tracking edges. *IEEE Transactions on Pattern Analysis and Machine Intellingence*, 26(4):479–494, April 2004.

[52] C. Stiller, J. Konrad. Estimating motion in image sequences (A tutorial on modeling and computation of 2D motion). IEEE Signal Processing Magazine, vol. 16, pp. 70-91, July 1999.

[53] W. B. Thompson, P. Lechleider, and E. R. Stuck. Detecting moving objects using the rigidity constraint. *IEEE Trans. Pattern Analysis and Machine Intelligence.*, 15(2):162–166, 1993.

[54] W. B. Thompson and T.-C. Pong. Detecting moving objects. *International Journal of Computer Vision*, 4:39–57, 1990.

[55] T. Veit, F. Cao and P. Bouthemy. Probabilistic parameter-free motion detection. *Proc. Conf. Computer Vision and Pattern Recognition, CVPR'04*, Washington DC, June 2004.

[56] J. Weickert, A. Bruhn, N. Papenberg, T. Brox. Variational optic flow computation: From continuous models to algorithms. International Workshop on Computer Vision and Image Analysis (ed. L. Alvarez), IWCVIA'03, Las Palmas de Gran Canaria, December 2003.

[57] N. Vasconcelos and A. Lippman. Statistical models of video structure for content analysis and characterization. *IEEE Trans. on IP*, 9(1):3–19, Jan. 2000.

[58] Y. Yacoob and J. Black. Parametrized modeling and recognition of activities. *Sixth IEEE Int. Conf. on Computer Vision*, Bombay, India, 1998.

[59] L. Wixson. Detecting salient motion by accumulating directionally-consistent flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):774–780, August 2000.

[60] L. Zelnik-Manor and M. Irani. Event-based video analysis. *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, Kauai, Hawaii, Dec. 2001.