Deliverable D15.3 MDS3

State of the Art of Content Analysis Tools for Video, Audio and Speech

DOCUMENT IDENTIFIER	PS_WP15_JRS_D15.3_SOA_Content_Analysis_v1.0
DATE	10/3/2005
ABSTRACT	This report surveys the state of the art of analysis algorithms and tools for audiovisual content and discusses their feasibility and practical usability, as well as the interdependencies between the various tools.
KEYWORDS	audiovisual content analysis, metadata extraction, video indexing, content abstraction
WORKPACKAGE / TASK	WP15
AUTHOR, COMPANY	Werner Bailer (JRS), Franz Höller (JRS), Alberto Messina (RAI), Daniele Airola (RAI), Peter Schallauer (JRS), Michael Hausenblas (JRS)
NATURE	Report
DISSEMINATION	Public
INTERNAL REVIEWERS	

DOCUMENT HISTORY

Release	Date	Reason of change	Status	Distribution
0.1	2004-08-10	Created	Living	Confidential
0.2	2004-11-24	JRS initial input	Living	Confidential
0.3	2004-12-10	Completed JRS input to 3-5 and 10	Living	Confidential
0.4	2004-12-22	Added input to part C	Living	Confidential
0.5	2005-02-04	RAI input to various paragraphs	Living	Confidential
0.6	2005-02-22	Part B and D&Q analysis	Living	Confidential
0.7	2005-02-28	Input on Video Content abstraction, Face Detection, VideoOCR	Living	Confidential
0.8	2005-03-04	Final draft	Living	Confidential
1.0	2005-03-10	Approved in WA MAD meeting	Final	Public

1 Contents

2 Document Scope

This deliverable is a report containing a survey of the state of the art of analysis algorithms and tools for audiovisual content. Audiovisual content analysis tools will be used in PrestoSpace WA MAD to automatically extract metadata from the essence for the purpose of:

- support manual annotation by content structuring
- provide input to semantic analysis tools
- indexing audiovisual content and make it searchable by text-based and content-based search methods.

3 Executive Summary

This report is a survey of the state of the art of tools and algorithm for analysis of audiovisual content for the purpose of metadata extraction. The report is divided in four parts: Parts A through C discuss content analysis tools working on visual, audio and jointly on audiovisual features respectively. Part D summarises the state of the art of the tools and discusses their feasibility and practical usability, as well as the interdependencies between the various tools.

Benchmarking results, such as those of TREC Video Retrieval Evaluation (TRECVID), are used to objectively evaluate the quality of audiovisual content analysis tools.

4 Overview

The task of audiovisual content analysis is to extract information from audiovisual material. The extracted information is metadata, i.e. data *about* the material, data *describing* the material. Content analysis can thus often seen as reversing the authoring process (cf. [Dim03][Snoek05]), as there an audiovisual material is produced from a information about the content to be produced (e.g. script, storyboard, scene drawings).

In [Snoek05], the authors define three basic questions of content analysis:

- what: the granularity to be described
- how: the modalities to be described
- which: the kind of index to be built

The modalities which can be used are the visual, audio and text modality. The first two are within the scope of this report. Concerning the text modality, only the extraction of text from the other modalities (i.e. video OCR and speech to text) will be covered, not the processing of text. The information extraction tools working on text and content analysis results will be will be discussed in the PrestoSpace deliverable MDS6 Semantic Interpretation Tools (D15.6).

Three perspectives for viewing audiovisual content have been proposed [Snoek05]: layout (e.g. shot structure, camera motion), content (e.g. people and objects appearing) and semantics (e.g. scenes, named events). Figure 1 illustrates these perspectives. These three perspectives also correspond to the levels of features that can be extracted from audiovisual content, ranging from low-level features describing layout to high level features describing semantics. A graphical overview of the interdependencies between the features is shown in Section 19.

There are three typical class of tasks involved in audiovisual content analysis. The first is video parsing or segmentation [Wang00a], i.e. decomposing an audiovisual item into temporal segments, which are characterised by some features. This starts with low-level features, such as visual or audio properties, for example, in the case of video, shot boundary detection has a special role as prerequisite for the following processing steps [Hanj04]. On a higher level, this includes parsing the video into semantic segments.

The next class are classification tasks [Wang00a], also often called indexing [Hanj04]. This includes for example the identification of genres, the detection of settings and the detection of people and objects appearing in the audiovisual content.

Finally, summarisation [Wang00a] or abstraction [Hanj04] deals with representing the content in a compact and structured way, highlighting the salient properties of the content.

We decided to organise this report according to the modalities for extraction the features. Parts A and B describe extraction tools which use only visual and audio features respectively, from low- to mid-level features. Part C describes tools which are jointly based on audiovisual features, and are thus capable of extracting higher level descriptions, very often using the results of the tools presented in parts A and B. Part D draws conclusions about the status and feasibility of the tools.

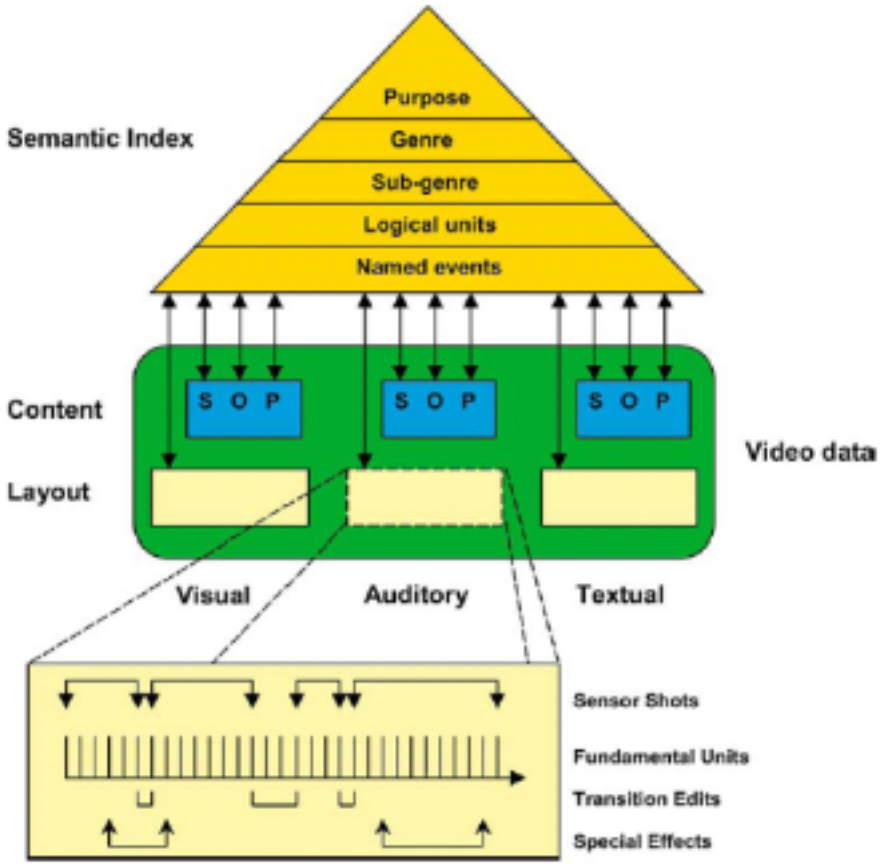


Figure 1: A unified framework for multimodal video indexing proposed in [Snoek05]. The letters S, O and P stand for setting, objects and people.

Part A: Visual Content Analysis Tools

5 Low-level Visual Features

One important goal of content analysis is to describe the media being analysed in a compact, efficient, exchangeable and transferable way. If such a description shall be unambiguous, and if it shall be possible to create such a description without further data than that of the media being analysed, one has to base the description on the features, which can be directly extracted from the media data.

Low-level visual features are single visual properties of a visual media, such as colour, texture, shape and motion. In [SchemaD21] they are described to be the “basic cues [of] our visual system”. Low-level features serve as a basis for a number of analysis algorithms for higher level features.

5.1 The Concept of Descriptors

As low-level features are typically contained by many or all pixels of an image (e.g. every pixel has a colour), it is necessary to derive more compact representations of these features. Such a representation is called *descriptor*, and represents a certain feature of an image, a set of images (e.g. an image sequence), or a region of an image. The definition of a descriptor shall consist of three parts:

- Representation: which data is contained in the descriptor and what is meant by it
- Extraction: how to create descriptor data from a given visual media
- Comparison: how to determine the distance or similarity between two descriptors of the same type

The first is of course mandatory, the other two may be optional. In that case, the descriptors are no longer exchangeable and may be only comparable if extracted and compared by the same application.

In the following the three parts are discussed in more detail.

5.1.1 Representation

The representation of a descriptor defines the organisation and meaning of the data of the descriptor. For low-level features, the data of the descriptor is usually a n-dimensional vector (or a set of vectors), where the number of dimensions depends of the feature space which is used. In the case of a single vector, each instance of the descriptor represents a point in the descriptor’s feature space.

The main design criteria for the representation are representativeness and compactness. The first requires that descriptors extracted from visually similar media are similar and that those extracted from visually different media are different, while the second requires that the descriptor has as little data as possible. Naturally, those two contradict one another, as the first would be best fulfilled by the whole media item, while the second would be best fulfilled by a single value. A further design criterion may be comparability, so that the descriptor can be efficiently compared and indexed.

5.1.2 Extraction

Extraction defines the process of creating a descriptor from a given visual media item. This process may include complex algorithms, but as discussed below, it is not as time critical as the comparison.

The extraction algorithm typically has to reduce the complexity of the data and to extract representative components. Thus, extraction algorithms often include clustering, component analysis or transform domain approaches.

5.1.3 Comparison (Matching)

The comparison method of a descriptor depends inherently on its representation. In the simplest case, the descriptor is a vector in an n-dimensional linear space, so that e.g. the Euclidian distance may be used. But for many common descriptors, e.g. histograms, this does not hold. A number of different metrics have been proposed in literature, an overview can be found in [SchemaD21].

As typically one or more query descriptors have to be compared with a large number of descriptors, efficient indexing structures should be used to avoid evaluating the comparison metric for each pair of descriptors, which may be time consuming. The choice of the comparison metric may be related to the availability of efficient index structures.

It has been stated before that efficiency in comparison is an important issue. A descriptor is typically only extracted once from each media items, but compared to others many times. Extraction is often done in a background process, while comparison is often done while a user waits for feedback.

Traditionally, literature about Content Based Information Retrieval systems makes a rough distinction between query-by-example and query-by-content application scenarios ([Dbim99]).

In addition to these typical in which, simplifying a bit, a set of descriptors extracted from an image acting as a query sample is compared with a database of image descriptors, comparison algorithms are also used to extract structuring information by means of data clustering techniques (see next paragraphs about audiovisual segmentation).

5.2 Descriptors for Visual Features

Since the 1990s, a growing number of descriptors for low-level visual descriptors has been proposed. Especially the research in content-based image retrieval (CBIR, cf. Section 5.3) has fostered the definition of new descriptors and comparison metrics.

For interoperability, standardised descriptor definitions are necessary, in order to be able to compare descriptors across application boundaries. The MPEG-7 standard defines a set a descriptors for low-level visual features. However, in order not to hinder further development, the part describing extraction and comparison of the descriptors is not normative, which of course weakens interoperability.

5.2.1 Colour

5.2.1.1 Feature Space

The feature space of a colour descriptor is a colour space, where each dimension in the space represents one colour component. A large number of colour spaces have been proposed in literature (cf. [Hunt87][Foley90]). These colour spaces have different properties, which influence the performance of the descriptors based on them. These properties include for example the correlation between the colour components and the capability to separately represent lightness and hue.

An important property is the distance measure that can be used in the colour space, as this determines the quality and meaningfulness of the matching result of two descriptors. Some colour spaces, for instance CIE Luv and CIE Lab, have been designed with the goal of perceptional uniformity. This means that the Euclidian distance between two colours in these colour spaces corresponds to the perceived distance between the two colours.

5.2.1.2 Classes of Colour Descriptors

It is not the intention of this work to list all the different approaches of colour descriptors that have been proposed in literature. The goal is to give an overview over the different approaches that have been used, and of their relative strengths and weaknesses. A discussion of a number of approaches for colour description and a bibliography can be found in [SchemaD21], an overview of the colour descriptors that are available in MPEG-7 is given in [Manj01].

Histogram based methods

The oldest and most common approaches use histograms to describe the colours of an image or an image region. Similarity is measured by histogram intersection. A number of different norms have been proposed for calculation of the intersection of two histograms. The MPEG-7 ScalableColor descriptor is an example for this approach, using the L2 norm for calculating the similarity between two descriptors.

Colour moments have been proposed as a reduced alternative to full histograms.

Histograms are space-independent representations and thus robust to displacement of parts in the image. One problem is that for practically calculating the histogram, the colour space must be quantised. Depending on the norm used for calculating intersection, this may overly decrease similarity of descriptors when colours are similar but not identical.

Dominant colours

To overcome the problem of quantisation caused by using histograms, describing a set of dominant colours of an image or a region has been proposed. These dominant colours can be described exactly and correspond to the salient colour features of the image or region. This kind of descriptors is more appropriate for regions then for whole images, as it is more likely to find a set of distinctive dominant colours in a region than in an image.

This kind of descriptors allows describing the salient colour features in an exact and space-independent way. Matching can be done in a feature space that has only the dimensions of a colour space, but it will require n^2 comparisons to match two descriptors with n dominant colours. In an image or larger region it may be difficult to find distinctive dominant colours.

Transform domain approaches

This kind of colour descriptors contains frequency components of a transformed colour image. The DCT is commonly used as transform, not only because it is real-valued, but also because the DCT coefficients are readily available in JPEG or MPEG-1/2/4 compressed image data. In the descriptor, the low-frequency components are used to describe the global colour features of a region or an image.

While transform domain colour descriptors are still scale and rotation independent, their AC coefficients contain some information about the colour distribution in the image or region. The extraction from arbitrarily shaped regions requires some additional steps as e.g. padding with the DC value.

5.2.1.3 Spatial information in colour descriptors

All of the colour descriptors discussed above do not contain information about the spatial location of the described colour features in the image. For many retrieval tasks it is desirable to have at least some basic information about the spatial location of different colours.

The simplest approach to considering location information is to split an image into blocks and extract a descriptor from each block. This of course causes problems at block boundaries, for which overlapping blocks have been proposed as a solution.

A further step consists in aggregating individual blocks into block regions (i.e. two-dimensional regions made up of composition of single blocks) on the basis of dominant colour matching, thus providing intermediate-level descriptions of colour disposition and colour spatial domain for images ([Mont04]). More complex approaches ([SD97]) use a content-driven segmentation to enhance the performance of descriptors comparison in the regions of the images that are deemed most important for a defined application.

Algorithms for aggregation of multi-modal subregions (i.e. regions of image not presenting a single colour dominant) have been proposed as well ([MM97][CRG97]). These approaches share the use of colour sets rather than of individual dominant colours to describe and subsequently compare image regions.

More advanced approaches try to model the local correlation of colours without fixing it to a certain location in order not to loose the properties of scale and rotation independency. These approaches include the colour correlogram approach and the MPEG-7 ColorStructure descriptor.

Other methods proposed in order to increase the performance of global image colour histograms without actually splitting the image into regions include the use of colour coherence vectors ([PZM96]) and spatial chromatic histograms ([CLPO99]).

Furthermore, as a general indication, some other features such as edge identification, may support generic image regionalisation with respect to the colour features.

5.2.2 Texture

Images or image regions retaining a well-recognisable and repetitive configuration of luminance values over the spatial dimensions are said to have a defined texture.

Quite intuitively, texture can be defined as a description of the structure of an image, with respect to its spatial- and frequency-related perceptive features. Periodicity, coarseness, directionality and contrast are common examples of descriptive dimensions of a texture descriptor.

Most texture descriptors work on grey level images only, some can be also applied to colour textures and some colour specific methods have been proposed. In the following, the most common classes of texture descriptors are discussed.

Obviously texture descriptors are most meaningful when applied to a homogenously textured region. To describe the textures in an whole image, and take the spatial properties of the different textures into account, the image is mostly split into blocks.

Again, as dually stated for what colour features are concerned, the outside use of colour and luminance descriptors may enhance image regionalisation of images with respect to texture features.

A discussion of texture descriptors in MPEG-7 can be found in [Manj01] and [Wu01].

5.2.2.1 Spatial Texture Features

Auto-Correlation

Structural features of images depend on the spatial dimension of their grey-level primitives. The auto-correlation function of an image is a measure of such dimension. In fact, coarse-grained images tend to have large primitives, in contrast with fine-grained images. According to these characteristics, the autocorrelation function decreases more rapidly or more slowly if the primitives are of small or large size respectively.

Co-occurence

The co-occurence matrix describes the correlation of spatially neighboured grey values and thus primitive texture features such as moments, correlation and entropy of a local neighbourhood can be derived from it.

Fractal analysis

Textures may be also described by means of the measure of their fractal dimension ([EzC02]). Typically, for natural images, the fractal dimension is related to the image roughness, i.e. an images with higher levels of roughness present higher fractal dimension.

Higher level feature descriptions

A drawback of the co-occurence approach is that not all of the features it describes are directly meaningful to humans. Therefore more meaningful features have been proposed for texture description. These include contrast, directionality, regularity.

5.2.2.2 Transform Domain Representations

Frequency Domain Approaches

Frequency domain approaches describe the energy of the components in certain frequency bands to characterise the texture. This allows for simple and efficient matching.

Wavelet-based approaches

A number of descriptors for texture use the Wavelet transform to extract the descriptor and describe the properties of wavelet subbands. The main advantage is that wavelet based texture descriptions allow for scale independent description of texture.

5.2.2.3 Texture signatures

In order to synthetically grasp and represent the features of a given texture, some authors ([Mich93][Hunt98][Piat00]) propose the use of texture signatures, i.e. of vectors of independently calculated texture features. The following signatures are the most commonly utilised.

Contrast, coarseness and directionality

Euclidean distances measured in this 3-dimensional vector space have been found to closely represent the distances perceived by humans.

Repetitiveness, directionality, complexity

Another representation of image texture is that obtained considering an image as a homogeneous 2-dimension discrete random field (Wold decomposition [LP94][LP96]). With this approach, an image is seen as composed of three independent contributions intuitively matching the named three visual characteristics.

Texture Energy

Following this approach, a set of energy images are calculated convolving original images with special functions named kernels ([Laws80]). Each original pixel is represented by a vector and texture signatures are obtained by applying operators in the resulting vector space.

5.2.3 Shape

Shape is an important feature of image regions or objects. Shape description and matching approaches are interrelated, as certain matching approaches require a certain kind of representation. More detailed versions of this kind of overviews can be found in [Sto04] and in [Mont04].

The shape description and analysis algorithms now proposed are, of course, processes that are applicable after that a spatial image segmentation is done. In general we can imagine, for the sake of an easy description, that the spatial segmentation provides image masks identifying each of the regions and objects in the images. Therefore, for example, boundary pixels for a certain region can be identified as those having at least an extra-region pixel as a neighbour.

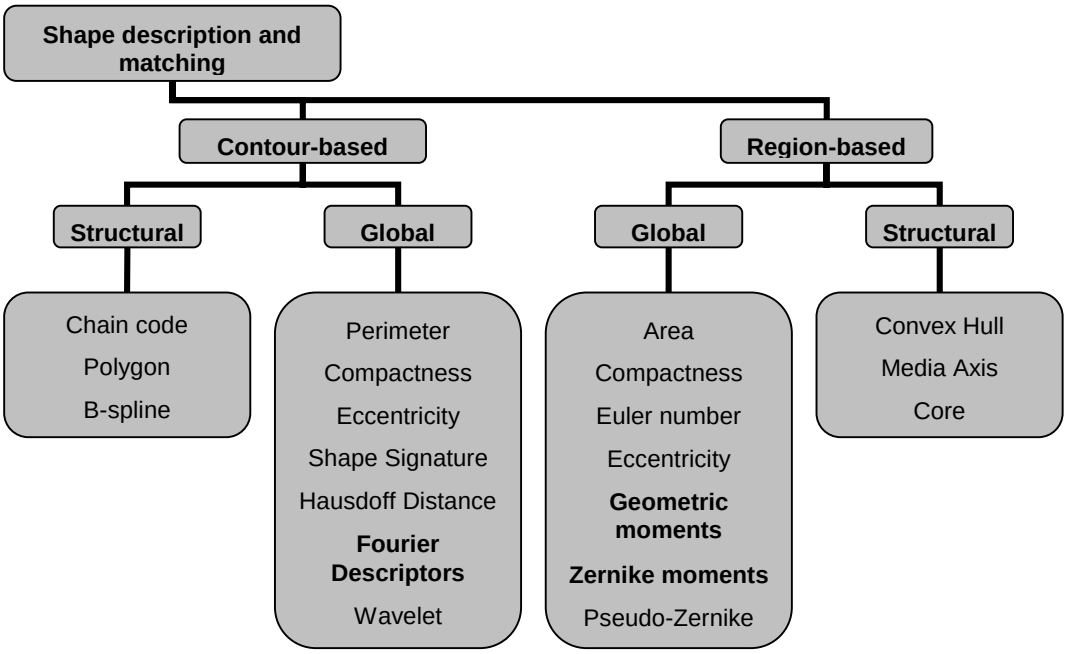


Figure 2: Overview of shape representation and matching methods.

5.2.3.1 Global Contour Based Methods

The simplest method to obtain a representation of a contour of a region is obviously to consider the coordinates of each pixel of the region boundary. But this is also the most memory-consuming method and the most prone to contour matching errors due to the high sensitivity to contour details. In order to limit the amount of needed information to be coded in a contour descriptor, it may be chosen to select from the entire set of boundary pixels only those that are deemed to be the "most important" from a perceptive point of view. These points are said to form the contour's interest set.

In general, a multi-dimensional numeric feature vector is computed from each shape boundary set.

A very simplistic approach for characterising contours is providing for each point of the interest set some basic features: the Cartesian coordinates, the local tangent direction, the radius coordinate and the curvature function. More complex approaches are those summarised in the following paragraphs.

The vectors obtained from these features are compared by using a metric distance, such as Euclidean or city block distance.

Shape Signature

A shape signature is any 1-D function representing 2-D boundaries, which usually uniquely describes the shape. Shape signatures have been studied extensively in [DAV97] and [OTT91].

There exist many shape signatures including *complex coordinates*, *polar coordinates*, *central distance*, *tangent angle*, *cumulative angle*, *curvature*, *area* and *chord-length*. Shape signatures reduce the matching problem from 2-D to 1-D, but shift matching is required. This method is normally sensitive to noise, so slight changes can cause large errors in matching. It is necessary to use further pre-processing before matching.

Fourier Descriptors

Fourier transform is a classical tool which has been used for many years in nearly every signal processing system and computer vision system. The fundamental principle behind Fourier transform is that a signal is represented by its spectral components. Different signals can be distinguished by the transformed spectra.

The Fourier transform can be applied to image, region or to region shape contour in a slightly modified form [DEN02].

Usually the Fourier descriptor is used for a derived shape signature, which is a 1-D shape border function (as described above). The result is a set of Fourier coefficients, which are a representation of the shape. There exist various ways to extract and compare Fourier shape descriptors. The number of coefficients needed to describe a shape sufficiently is about 10 in the best Fourier approaches [DEN02].

Wavelet Descriptors

This is another spectral description approach which is similar to Fourier descriptors (FD). Both of them are usually derived from 1-D shape signatures. Wavelet descriptors have the advantage over FD in that they provide spatial multi-resolution, but the increase of spatial resolution sacrifices frequency resolution. The necessity of performing shift matching when comparing Wavelet descriptors reduces its practical usability. Matching also depends on the shape complexity. [YAN98] contains a detailed discussion of Wavelet descriptors.

Contour distributions

Contour distribution is a 2-D function representing the distribution of radius and angle of a contour with respect to a chosen origin of coordinates. Contour distribution is rotation-invariant (radius) and scaling-invariant (angle). A contour distribution vector suitable for content retrieval and image matching applications can be obtained by discretisation of each of the two components of the distribution.

Scale Space

Before explanation of scale space is necessary to define the term inflection point. A point on the curve, where a convex part verges on a concave part or vice versa, is called an inflection point. Second derivative of a curve in the inflection point equals zero.

The scale space representation [DEN02] is created by tracking the position of inflection points in shape boundaries filtered using low-pass Gaussian filter of variable widths. The inflection points that remain present (the most important inflection points) are assumed to represent significant object characteristics. The result is a so called fingerprint consisting of inflection points. This approach is also known and studied as Scale Space Filtering ([Flor00]).

Curvature Scale Space (CSS)

Basically, CSS processes shape boundaries as a 1-D signal. By examining inflection points, it finds convexities and concavities of the boundary.

The process of extraction of a CSS descriptor is described in the following. The first step is scale normalization which samples a fixed number of points from the shape boundary. Then the curvature is computed. Curvature zero-crossing points are then located and written to the CSS map. The shape is then evolved into next scale by applying Gaussian smoothing. The wider the range of filter, the smoother is the resulting shape. New curvature zero-crossing points are then located at each evolving scale. This process ends when there are no zero-crossing points in the curvature. The position of each inflection point is then written to the CSS map. At the next step the peaks in CSS map are found and used as a CSS descriptor.

A disadvantage of this descriptor is that it is not robust in a global sense [DEN02]. Due to this problem, additional global descriptors like eccentricity and circularity are used with CSS.

In the following, matching of CSS descriptors is discussed. Rotation of the object causes circular shifting of CSS peaks in CSS map [DEN02]. Mirroring of the object causes mirroring of the CSS peaks. Because of this, CSS matching algorithm has to shift the one descriptor to overlay the highest peak to the highest peak of the other descriptor. For robustness reasons, the two highest peaks of each descriptor are used, and so there are four combinations to be evaluated. To fix the mirroring problem, the compared shape CSS has to be mirrored and compared again. So each descriptor matches results in 8 basic comparison operations.

The MPEG-7 ContourShape descriptor [MPEG7-3] is based on CSS.

Autoregressive Method (Stochastic Method)

Autoregressive approaches [SEK92] are based on stochastic modelling of 1-D shape contour functions. Linear autoregressive models express a value of a function as a linear combination of a certain number of preceding values.

The disadvantage of this approach is that in the case of complex boundaries a small number of parameters are not sufficient for description.

5.2.3.2 Structural Contour Based Methods

In structural approaches, shapes are broken down into boundary segments called *primitives or tokens*. Structural methods differ in the selection of primitives and their organisation for shape representation. The result is usually encoded into a string.

Chain code

Chain codes describe an object as a sequence of unit-size line segments with a given orientation. In 1961 Freeman [FRE61] described a method permitting the encoding of arbitrary geometric configurations. In this approach, an arbitrary curve is represented by a sequence of unit length vectors with a limited set of possible directions. Further steps are necessary to make the chain code starting point independent.

It is not advisable to use chain code for object matching, because it is very sensitive to boundary noise and variations and the dimension of this code is high [ZEL03].

Polygon

The polygon approach [GRO92] is based on breaking down the shape boundary into line segments by polygon approximation. Then the polygon vertices are used as primitives. The features for each primitive are expressed as a four element string which consists of internal angle, distance from next vertex, and the vertex' x and y coordinates. Obviously this representation is not rotation invariant. Similarity between two features is measured as the *editing distance* of two feature strings.

Polygon representation can be used for man made objects retrieval, but the application of it for general shapes is impractical.

Curvature and orientation

In this approach ([BDbP99]), the shape contour is split in tokens defined as the segment of the contour occurring between two consecutive relative minimum or maximum points. For each token, a bidimensional vector is calculated representing its maximum value of curvature and its orientation. A contour is then represented by a dynamic-length list of such vectors.

Syntactic Analysis

Syntactic analysis [DEN02] [ZEL03] is inspired by the idea that composition of a natural scene is analogue to the composition of language. It means that sentences are made from words, words from letters and that there exist rules for creating the sentences. Syntactic analysis uses a set of primitives. For matching string matching approaches are used, which find the minimal number of edit operations to convert one string into another. It is also possible to convert similar shapes to grammar classes (different languages) and then make a matching by deciding in which language the shape representation makes a valid sentence.

This approach is not practical for general use, because it is impossible to infer a pattern grammar which can generate only valid patterns.

Shape Invariants

Like in syntactic analysis, the shape is represented by boundary primitives (so called *invariants*) in this approach. This technique attempts [SON93] to represent properties of boundary configurations, which remain unchanged under an appropriate class or transform. Invariant theory is based on a collection of

transforms, which can be composed and inverted. If an object cannot be represented by lines or curves, differential invariants can be formed.

This approach has several problems. Invariants are usually derived from pure geometric transform of shape, but in reality shape rarely change according to strict geometric transform, foremost the shape of non-rigid objects. Furthermore it needs some form of sub-graph matching, which is a NP-complete problem.

5.2.3.3 Global Region Based Methods

Simple geometric properties

Also in the case of region characterisation, some basic geometric features can be used to synthetically represent the part of a region or object enclosed by its contour. Examples are: region area, Euler number (also known as genus, a topological property of regions), orthogonal projections, eccentricity, elongation, compactness. Similarity matching methods based on these simple features doesn't usually present a good resolution an precision.

More accurate developments split a single region into smaller regions that are considered standard basic blocks (e.g. ellipses and rectangles).

Geometric Moments

There is a lot of new literature about geometric moments, which further extends the work originally described in [HUM62]. This approach is based on the work of the 19th century mathematicians Boole, Cayley and Sylvester, and on the theory of algebraic forms.

A set of moment invariants (usually called *geometric moment*) is derived from using nonlinear combinations of the lower order moments. The features are invariant to rotation, scale and translation. The main problem with this approach is that only few lower order moments are not sufficient to accurately describe the shape and it is difficult to derive higher order invariants.

Algebraic Moment Invariants

Algebraic Moment Invariants have been described in [TAU92]. For representation of the shape take the first *m* central moments and are used as eigenvalues of predefined matrices, whose elements are scaled factors of the central moments. This method tends to work well on objects, where the distribution of pixels is more important than the shape.

Orthogonal and Other Moments

Teague introduced orthogonal moments in [TEA80]. Legendre and Zernike moments originate by defining Legendre or Zernike polynomials with orthogonal basis. These moments are so called *orthogonal moments*. Other orthogonal moments are pseudo-Zernike moments which use real-valued radial polynomials in Zernike polynomials as the model transform kernel. Furthermore non-orthogonal moments like *complex moments*, *rotation moments* and *geometric moments* can be used.

Results show that geometric moments, complex moments and pseudo-Zernike moments are less affected by noise, while Legendre ones are severely affected by noise. Among the different moment shape descriptors, the Zernike moments are the most useful. Because they are extracted in the spectral domain, they have similar properties as other spectral features which are well understood.

The disadvantage with many moment methods is that it is difficult to correlate high order moments with shape physical features.

Angular radial transform

The angular radial transform approach is a Copley transform defined on an image in polar coordinates. The basis functions are separable in angular and radial directions. The texture is described by the coefficients of the basis functions.

The MPEG-7 RegionShape descriptor [MPEG7-3] is based on the ART.

Grid Method

Basically, a grid of a certain number of cells is overlaid on a shape. The grid is then scanned from left to right and top to bottom. The cells covered by shape are assigned 1 and the uncovered cells are assigned 0.

The result is a bitmap which is then represented as a binary vector. For similarity measure the *Hamming distance* or *city block distance* are used. In order to accommodate translation, rotation and scaling of the shape, the shape is first normalised.

The advantage of this method is its simplicity in representation. It is also well understood. However, the computation is expensive.

Shape Matrix

The idea of this method is normal raster sampling. It has been proposed by Goshtasby [GOS85]. In contrast to other sampling methods, the shape matrix method uses a polar raster. The matrix is than formed so that columns correspond to circles and rows correspond to lines. The grid is overlaid in the center of shape. To make shape matrix rotation invariant, it is necessary to normalize the shape at first.

5.2.3.4 Structural Region Based Methods

Convex Hull [SON93]

A region R is convex if the line created from any two points is wholly inside the region R. The convex hull H of a region is the smallest convex region which contains all the points of the contour. Because of shape boundary noise, the boundary function is smoothed as a pre-processing step. The shape is represented by a string or tree of concavities.

Each concavity can be described by its area, bridge length, maximum curvature, and distance from maximum curvature point to the bridge. The matching is done by the string or graph matching.

Medial Axis

The idea of this approach, originally called region skeleton ([DEN02][ZEL03]), is to store only the topological information concerning the structure of object. The medial axis is the locus of centres of maximal disks that fit within the shape and touches the shape in at least two points.

The skeleton can be decomposed into segments and represented as a graph according to certain criteria. The matching then becomes a graph matching problem. The computation of medial axis still remains a challenging problem.

5.2.4 Motion

The main goal of identifying features related is to describe the movements of the objects taking part in the overall shooting event. The movements of either of these types of objects produce characteristic changes in the audiovisual recorded content. The interesting objects include the actually shot objects (or *subjects*, e.g. persons, mountains, buildings) and the shooting cameras (*sensors*).

The fundamental assumption underlying any of the methods that will be described in the following sections is that relevant motion of real objects during shooting (i.e. subjects and sensors) can be inferred, with different levels of precision, by the analysis of the three-dimensional signal represented by the recorded video sequence.

Motion is also used as an input feature to infer the presence of, to identify and to track objects over the sequence (see Section 6.2).

There are basically two types of motion that can be described in an image sequence: global motion, i.e. the motion to which all pixels in the image are subject to, and object motion, which describes the relative motion of regions to one another or to a background object.

In this section, we review approaches for motion estimation and ways for the description of motion. Motion based spatio-temporal segmentation is discussed in Section 6.2.

5.2.4.1 Motion Estimation Approaches

A great variety of methods for solving the motion estimation problem has been proposed. Stiller and Konrad define in [Still99] three basic elements of algorithms for estimating motion trajectories: a *motion model*, an *estimation criterion* and a *search strategy*. Motion estimation methods could be classified by these three elements, but many algorithms are similar in some of the elements. In the section we will not discuss all these elements in detail, but only give a brief overview about motion models (a more detailed review can be found in [Bai02]). Then we discuss the general approach of hierarchical motion estimation and three classes of motion estimation algorithms, following the classification in [Tek95].

Motion Models

The motion model defines the number of parameters of the motion, and thus the allowed degrees of freedom and the complexity of the motion description. The motion estimation problem is underconstrained, i.e. it is not solvable unambiguously from the input data. A motion model is necessary, as it incorporating further assumptions about the motion to be estimated. Depending on the region of support (the area which is subject to the motion being estimated) and knowledge or assumptions about the motion's structure, one has to select an appropriately complex motion model.

The simplest and most flexible motion model is the *translational model*, which allows each point to move independently by a displacement vector in x and y direction. This model does not constrain the motion vector field in any way, but the estimation requires local two-dimensional image structure.

Therefore, additional constraints, e. g. a spatial smoothness constraint, that allow to define a motion vector influenced by its neighbours in sparsely structured regions, are used. Because this model allows each point to move individually and without any further constraints in its simple form, it is also referred to as a *non-parametric model* [Anan93].

Models that include more parameters (often called *fully parametric models*) can only be used with a sufficiently large region of support, as it will otherwise be impossible to estimate the parameters. They constrain the freedom of motion by assuming that all points within the region of support move uniformly. On the other hand, the parameters can be estimated highly accurately, as all points in the region contribute to the result. The motion of a planar surface under orthographic projection can be described using a six parameter *affine model.*

For surfaces which are distant enough from the camera and not too large, this model also gives a good description in the case of perspective projection [Anan93]. In many cases, also a reduced four parameter version of the affine model, describing translation, rotation and scale will suffice. The reduced model is not capable of modelling shear and different scale factors along the axes, but it is more stable because of the smaller number of parameters to be estimated.

For fully describing a planar surface under perspective projection, an eight parameter model, called *projective linear*, is required.

A more comprehensive overview of commonly used motion models, their formal descriptions and capabilities can be found in [Still99].

A specific model for the estimation of camera motion has been described in [Tan00]. It is based on a six parameter projective model without translation, in which the six parameters depend on the initial focal length in a sequence, the change factor of the focal length and the rotations of the camera around the three axes in space.

Hierarchical Motion Estimation

Hierarchical or multiresolution motion estimation refers to motion estimation methods that work successively on instances of the original image with different resolution. The stack of images with different resolutions originating from one image is called image pyramid and the images' sampling lattices usually differ by a factor of two. The pyramid can be generated by low-pass filtering and subsampling of the original image [Still99].

Hierarchical motion estimation can be used with any motion model, estimation criterion and search strategy. The hierarchical approach offers two benefits: Firstly, it makes it easier to estimate large

motion, as large motion will appear as smaller motion on higher levels and image details and noise will disappear. Some methods, especially those that require a search area, are unable to find large motion without a hierarchical approach. Secondly, the computational cost for motion estimation on higher levels is small because of reduced image data and limited search areas, while a large amount of the motion in the image can be found on these levels.

The motion estimation is usually done from the coarsest (highest) to the finest (lowest) level. The estimate from the previous level has to be projected to the next level and serves as initial estimate for the motion estimation on this level [Anan93].

Optical Flow Methods

This section discusses motion estimation methods based on extensions of the optical flow equation. There are both methods using non-parametric and parametric motion models. The aim is always to add additional constraints for the motion vector field or to enlarge the region of support in order to overcome the underconstrainedness of the motion estimation problem.

Minimizing the regularisation criterion is a way to calculate a dense motion vector field. This method, as well as others presented here, requires the estimation of the spatial and temporal image gradients. This can be done by local polynomial fitting [Tek95], which requires the estimation of coefficients and therefore more computation. A simpler method, averaging four finite differences, has been proposed by Horn and Schunk [Horn81].

Lucas and Kanade [Luc81] proposed a method that applies the optical flow equation to a block instead of a single pixel. The error of the optical flow is then summed over all pixels of the block. A weighted summation can be used to decrease the influence of pixels at the borders of the block. This method imposes a strict constraint on the motion vector field and allows only translational motion of a block [Tek95].

If we consider larger regions of interest, the optical flow equation can also be used to estimate the parameters of a parametric, e.g. an affine motion model.

Block Based Methods

Block based motion estimation methods split the image into a number of blocks and assume the motion in each block to be uniform. These methods are widely used because of their simplicity, e.g. in compression standards like H.261 or MPEG 1 and 2. Most of these methods use translational motion models, which cause the estimation to fail for any kind of non-translational motion of a block. Block based methods are not restricted to a certain motion model, they can also be used with fully parametric models. The motion parameters are always applied to a whole block, which may result in serious blocking defects, as adjacent blocks can have different motion parameters, but the block border will usually not coincide with an object border. As a solution, overlapping blocks have been proposed, where the pixels in the overlapping area are assigned the average of the motion parameters of the blocks they belong to, or one of the possible motion parameter sets by hypothesis testing.

Any of the estimation criteria discussed above could be used in block based motion estimation methods. Block matching methods search within a certain area, the search area, for the best matching block, which was transformed according to the motion model (e.g. translated or affine transformed). To determine how well two blocks match, a matching criterion has to be defined; the mean square error (MSE), the minimum absolute difference (MAD) and the cross correlation are common choices. A number of different search procedures have been proposed to avoid exhaustive search of the search area. Most block matching algorithms work hierarchically in order to be able to estimate large motion while keeping search areas small.

The frequency domain phase correlation criterion has some desirable properties for use in block based methods. The region of support for this criterion should be rectangular because of the calculation of the DFT, it does not require a search strategy and it is able to find multiple motion in a block in order of reliability, so that it is not necessary to assign the same motion parameters to all pixels in a block, which avoids blocking defects [Tek95].

Bayesian Methods

Bayesian methods are based on a Bayesian estimation criterion (cf. [Still99]). Different algorithms are mainly characterised by the choice of structural, observation and especially motion models and the optimisation method.

The structural model relates the estimated motion vector field and the underlying image data by the effects caused by motion, e.g. luminance intensity changes, contrast distribution changes. It therefore serves as a measure how well the estimated motion fits the image data. The structural model is expressed as the conditional probability, that the predicted realisation of the random field describing the next frame equals the actual next frame, given the estimated value for the motion prior model, i.e. the motion vector field.

The structural model can be extended by an observation model describing the image acquisition process, e.g. filtering or addition of noise [Dub93]. Often the properties of the image acquisition process are not well known and so, for example, the zero-mean white Gaussian is a common choice [Still99].

Motion models express prior knowledge or assumptions about the motion vector field. The most common motion model is a formulation of the smoothness constraint, i.e. the assumption, that adjacent pixels will have similar motion.

The spatial smoothness model presented above has the same drawbacks as the smoothness constraint in proposed in [Horn81]: it smoothes across motion boundaries and it has problems with occlusions. One possible solution is the introduction of a line field modelling boundaries between areas of uniform motion, which is a binary Markov random field, based on a sampling lattice that is shifted a half pixel vertically and horizontally against the image's sampling lattice.

Line fields may become instable and cause problems with occlusion areas [Wei97b]. As an alternative, region labelling has been proposed. Instead of modelling a separate line field, a Markov random field containing region labels, based on the same sampling lattice as the motion vector field, is introduced. Sites having the same label are subject to the same motion and therefore the smoothness model may be applied.

Explicit modelling of occlusion (i.e. covering and uncovering of regions because of object motion) fields has been proposed. The occlusion field is then defined as a Markov random field on the same sampling lattice as the motion vector field. Each site has one of the states “covered”, “uncovered” or “moving/stationary”. The smoothness constraint is not applied to occluded pixels, the occlusion state is assigned based on a threshold for not finding a match in one adjacent frame and using the assumption, that the boundaries of occlusion regions are modelled in the line field or coincide with boundaries in the region label field [Dub93].

All Bayesian methods depend crucially on optimisation methods that are capable of finding the minimum energy state of the random fields and so the maximum a posteriori probability. All the techniques discussed here use a visiting scheme defining the order to update the sites. The order does matter, as the potential functions used for evaluating the total energy are based on adjacent sites, that may have been updated or not. A common visiting scheme is therefore a checkerboard scheme, which has the property that in the first pass half of the sites are updated based on the neighbour's previous state, while in the second pass the 4-neighborhood of each site has already been updated [Win95].

Common optimisation methods are Simulated Annealing, Iterated Conditional Modes (ICM) and Highest Confidence First (HCF), of which the first is stochastic and the latter two are deterministic.

5.2.4.2 Motion Description

In this section we discuss way to describe motion in a reduced and abstract way, i.e. beyond describing motion as the field of motion vectors for every pixel in the image (a so-called *dense* motion vector field). Motion description can be regarded as an information extraction process that produces concise descriptions of what is happening in the scene from the point of view of motion.

Camera Motion

Camera motion is usually a very interesting information in applications where users are searching for succinct connotations of the syntactical characteristics of a scene from an editorial point of view.

It is possible to estimate the dominant motion to which the whole image is subject to. However, this is often but not necessarily the camera motion. For example, if a large object moves in front of the camera, the dominant motion will be estimated as the dominant motion of the image, even if the camera was static. If camera motion is determined analytically from an image sequence, it cannot be fully reconstructed, as it is not possible to discriminate between different types of camera motion that cause the same visual effect (e.g. pan left and track left, if the target that has been shot is distant and the amount of motion is small, the translation of the camera cannot be distinguished from a rotation around its vertical axis).

Many approaches to camera motion estimation ignore the fact that camera motion can only be determined reliably over larger time range and accept the most dominant motion between a frame pair as the camera motion. The alternative is to estimate a number of dominant motions and then decide over a longer time range. For the estimation of multiple dominant motions clustering methods can be used, also the use of the RANSAC algorithm has been proposed [Chan04].

The selection of a dominant motion is based on the assumption that the camera motion is the most dominant one (i.e. those with the largest region of support), and that it is consistent and smooth over the time range.

The estimated camera motion can be described using one of the motion models discussed above. However, the description is more intuitive if the motion is described with common terms for camera operations, such as pan left/right and the amount of motion, even if not all of them can be determined analytically from the image sequence. The MPEG-7 camera motion descriptor [MPEG7-3] allows to describe camera motion with a relative fraction and an amount with the terms pan left/right, tilt up/down, roll clockwise/anti-clockwise, track right/left, boom up/down, dolly forward/backward.

Motion Trajectories

A motion trajectory is the description of the motion path of an object over time. This requires of course segmentation based on motion (cf. Section 6.2).

The information can be used to infer higher level information (e.g. relative motion of objects towards the background or one another) or can be used as a property for retrieval.

Motion Activity

In some cases it is not necessary to exactly estimate the dominant motion or the motion of the objects in a scene, but just to get a measure of the amount of motion in the scene, also called motion activity or scene dynamics. This information can for example be used for shot boundary detection, for shot classification or scene segmentation.

Depending on the application, the amount of motion, the dominant direction of motion and the spatial and temporal distribution of motion within a segment of an image sequence can be described (cf. MPEG-7 motion activity descriptor [MPEG7-3]).

5.3 Use of Low-level Visual Descriptors

The extracted low-level visual descriptors can be used either as an input for the extraction of higher level information (such as structuring or classification) or they can be used directly for similarity matching based on these descriptors.

5.3.1 Extraction of higher level information

We will not go into detail here about the algorithms that benefit from low-level visual features, but just present common classes of these algorithms. Whenever higher level analysis algorithms discussed in this report are based on low-level visual information, it will be stated there.

A class of higher level analysis algorithms that rely on low level visual features are segmentation algorithms. Segmentation can be done spatially, i.e. an image is decomposed into regions which are characterised by uniformity of a low-level visual feature such as colour, texture or motion, or temporally. Temporal video segmentation usually starts with shot boundary detection and goes to segmentations on a semantic level such as scene segmentation. Spatial and spatio-temporal segmentation are discussed in Section 6, shot boundary detection in Section 7.

Media structuring techniques aim at creating hierarchical structures whose elementary units are temporally contiguous parts of a media, thus reconstructing a possible semantic decomposition of content. Media structuring can be attempted using low-level features extracted from audio and video by means of cross-similarity calculations among different temporal intervals chosen e.g. downstream of a shot detection process (thus creating shot aggregates). In particular application environments (newscast programmes structuring), these methods show a good level of precision and recall given by the relatively simple and repetitive format presented by these contents. Various approaches have been proposed in literature for the solution of media structuring based on the analysis of low-level features, among which [YYL96]. Scene and story segmentation approaches are discussed in Section 14, media content abstraction techniques in Section 17.

Another class is represented by classification algorithms, which assign class labels to images or shots based on their visual content. For example, texture descriptors can be used to discriminate between natural and man-made structures in visual content (for example [Naph03]). Approaches to shot and scene classification are discussed in Section 15.

5.3.2 Similarity matching

A big part of research in content-based image and video retrieval has to do with similarity matching. The basic idea is that the user starts with a query example and retrieves a set of visually similar media items ([Dbim99] describes a comprehensive set of such systems). Depending on the types of descriptors used, a sketch may be sufficient as query example.

The two main problems of retrieval based on similarity search are the following: the user needs a query sample, which is close to what he expects to find. Such a sample may not always be available when beginning a query. This can be facilitated by providing a set of common example images, which the user can use. The second is the correlation between what is similar in terms of descriptors and what humans perceive to be similar, as human perception is not reduced to low-level visual features but takes semantic information into account when judging similarity. Thus similarity matching should be used in connection with other query option, such as textual or conceptual queries. In many cases similarity matching will not be the primary query option, but used for refinement or sorting of the search result.

An overview of content-based image retrieval systems, which use similarity matching based on example or sketch queries, can be found in [Velt02].

6 Spatial/Spatiotemporal Segmentation

Segmentation is the decomposition of an image or an image sequence into parts which are characterised by uniformity of a certain feature or a group of features. Some of the low-level visual features discussed in Section 5 can serve as features for segmentation.

Colour and texture are features used to segment still images or single frames of an image sequence into regions which homogenous colour or texture (spatial segmentation or intraframe segmentation). This is a prerequisite for extracting meaningful shape descriptions of regions.

Motion is the key feature for spatiotemporal segmentation (interframe segmentation), as it delivers regions which are subject to the same motion. Tracking these regions throughout a sequence results in segmentation of moving regions.

Purely temporal segmentation can be also achieved by the analysis of the temporal variation of low level features (e.g. a shot detection tool can be based on the variation in time of the average luminance content of images). Section 7 discusses approaches to shot boundary detection.

6.1 Spatial Segmentation

6.1.1 Approaches to Spatial Segmentation

6.1.1.1 Thresholding

Thresholding is the simplest segmentation approach, which makes decisions based on local pixel information. Simplest approaches just analyse the histogram, trying to determine signal peaks. While global thresholding uses the same threshold for all image pixels, adaptive techniques change the threshold dynamically over the image.

6.1.1.2 Edge Based Segmentation

Edge based segmentation approaches concentrate on contours and exploit spatial information, using filter kernels like Sobel or Prewitt. The weakness of this approach is that blurred and soft edges lead to broken contours. Thus it is not possible to work on image with lower resolution, as edges will be softened during downsampling.

6.1.1.3 Active Contour Model - Snakes

The active contour model is a more recent approach and is sometimes denoted as the Snake concept [Lee02]. Originally developed by Kass et al. ([KWT98]), the main idea of this approach is to represent boundary shapes as spline curves and iteratively modify the curve according to a special energy function. This function can be a gradient vector flow or e.g. the new concept of a virtual electric field.

6.1.1.4 Texture Segmentation

Segmentation can be seen as splitting the image into homogenous regions. The main difficulty for any texture segmentation approach is to define homogeneity of textures in mathematical terms. The main concept for texture segmentation is to model the textured image as an instance of a random field. Markov-Random-Fields, Gibbs and auto-regressive random fields are commonly used [Lu98]. A simpler approach considers comparisons of image sub-regions (e.g. blocks) in terms of their texture signatures ([Mont04], see also 5.2.2.3) and aggregation of blocks sharing the same dominant signature.

6.1.1.5 Normalised Cut

The normalised cut approach to segmentation was proposed in 1997 by Shi and Malik [Shi00]. It is based on a graph partitioning problem, where the nodes of the graph are the entities to become partitioned and the edges represent the strength with which the nodes belong to the same group. The target is to partition the graph, so that similarity within groups is large and similarity between groups is small.

The normalised cut approach can be applied to intra- and interframe segmentation. In the case of intraframe segmentation, the nodes are the pixels of the image and the edges correspond to how much two pixel match in intensity, colour, texture. In case field of interframe segmentation the nodes are space time triplets, while the edges describe motion similarity.

6.1.1.6 Object Detection

Another facet of the problem of segmentation consists in the individuation of new objects coming into the scene. In fact, if regions could be detected and classified as novel as soon as they appear, their influence on the already detected regions can be isolated and rejected, thus enhancing the performance of the overall motion estimation algorithm. Recently an interesting approach has been presented that is based on the use of neural networks for the novelty detection in video sequences ([SiMa04]). The aim of this work is to recognize natural objects in video sequences using adaptive network configuration to isolate new classes of objects during the playing of video content. Spatial image segmentation is achieved using a region growing approach. The drawbacks of this kind of

approaches lies in the fact that a training phase is always needed, that may make the algorithm sensitive to specific classes of content.

6.1.2 Colour Segmentation

Colour is an important and the most straight-forward feature that humans perceive when viewing an image. Thus colour is one of the most important candidates for intraframe segmentation.

6.1.2.1 JSEG – Colour Image Segmentation

JSEG is a fully automatic colour image segmentation scheme, which separates the segmentation process into two independent steps: colour quantisation and spatial segmentation [Den99].

6.1.2.2 Mean-Shift Algorithm for Colour Segmentation

Mean shift segmentation [Com97] is a feature space analysis concept, with the aim to estimate cluster centres in a specific feature space. The algorithm can easily be applied to colour segmentation. The pixels are mapped into the colour space (the feature space) and are clustered, with each cluster representing a homogenous region in the image.

The main strategy of the algorithm is the movement of a search window towards the cluster centre, whereas the radius of the window is a parameter of the algorithm. After choosing an initial position for the search window, it is iteratively moved to the desired position by calculating the so called mean shift vector and translating the search window by that amount.

The intensity distribution of each colour component can be viewed as a probability density function. The mean shift vector is the difference between the local mean of this function and the geometrical centre of the window and is proportional to the gradient of the probability density at the specific point. Thus the algorithm terminates when it reaches the highest density region of the probability density function, which represents the desired cluster center.

6.1.2.3 Morphological Greyscale Analysis

The field of mathematical morphology provides a wide range of operators commonly used in image processing. For image segmentation it is necessary to detect object boundaries in order to split the image into homogenous regions. A common approach for detecting object boundaries is to locate high greyscale differences and greyscale operators are applied to enhance these variations [Vin93].

The gradient is a suitable approach to detect greyscale differences. It is very important to filter the image before applying the gradient operator to reduce the sensitivity to noise. The morphological gradient is different to usual gradient operators. It is defined as the difference between the results of dilation and erosion operation. It is capable of generating sharp grey level transitions.

Applying morphology to colour images needs further definition and assumptions, because of the lack of a natural sort order of multivariate data and the mathematical differences due to the choice of different colour spaces. Therefore the extension of the concept of morphology to colour images crucially depends on the definition of an appropriate ordering of the colours in a certain colour space. Perceptually uniform colour spaces are preferred for this purpose, as the perceptual colour distance corresponds to the Euclidian distance of colours.

6.1.2.4 Watershed Segmentation

Watershed segmentation is based on the idea that every greyscale image can be interpreted as a topographic surface. The black pixels correspond to the ground of the surface white pixels represent the peaks of the surface. This topographic surface can then be theoretically flooded starting at regional minima. Thus, the image is separated in its catchment basins which are separated by so called watershed-lines. If this algorithm is applied to a gradient image, the catchment basins correspond to homogenous grey level regions of the image [Beu91], [Vin91].

In practice the watershed transform produces an over-segmentation due to noise or local irregularities. Thus several approaches for improved watershed segmentation exist. A commonly used concept to avoid over-segmentation is to use pre-defined markers, whereas the topographic surface is flooded

form these pre-defined points (and not the local minima). This approach is based on the assumption that vision systems often roughly know the location of the objects to be segmented.

6.2 Motion Segmentation

One of the most suitable features for segmentation is motion. Segmentation of regions is a crucial step for object recognition. While an object can consist of several differently coloured or textured regions, the whole object is usually subject to the same motion in natural scenes. This makes motion segmentation capable of segmenting semantically meaningful regions, provided that they are not stationary.

Basically, there are two ways to face the problem: either starting with motion estimation, and then segmenting the resulting motion vector field, or by segmenting each frame using a spatial feature and then estimating the motion for each of the regions found.

Motion segmentation, also referred to as spatio-temporal segmentation, is a “chicken and egg problem", as accurate motion estimation, especially of motion boundaries, relies on segmentation. Therefore, motion estimation and segmentation should be done simultaneously [Duf96].

Most motion segmentation methods aim at finding disjoint regions, so that all points in a region are subject to the same motion and no adjacent regions have the same motion. Wang and Adelson [Wa93] propose to represent each region by a layer which is only opaque where the moving object is present. This allows the representation of transparent motion and avoids explicit occlusion areas, while the motion vector field on each layer may be smooth.

6.2.1 Segmentation Based on Motion Vector Fields

Clearly this approach requires the estimation of a motion vector field as first step, for which a non-parametric (translational) motion model must be used because of the lack of region information. Then parametric (e.g. affine) motion models are fit to the motion vector field by starting with an initial segmentation (either a block based one or using the segmentation result from the previous frame) and estimating motion parameters for each region. Pixels are then clustered by the displaced pixel difference (DPD) with respect to the motion models and finally segmentation is refined [Duf96].

The problem with this approach lies in the estimation of the motion vector field, which is very likely to lack accurate motion boundaries and may contain falsely matched occlusion regions.

6.2.2 Motion Estimation Based on Segmentation

The problem with motion segmentation is that the segmentation criterion, i.e. the motion, is not known a priori. Instead of motion, another segmentation criterion is used in some algorithms to create an initial segmentation. Motion can then be estimated for the regions of the initial segmentation and used for refining it.

A simple way to initially segment the image is to split it into a number of blocks and thus evaluate possible block aggregations by means of block features similarity (a general approach can be found in [Mont04]). The initial segmentation can also be found by quantifying the greyscale levels in the image and define regions with similar greyscale values, though this may lead to oversegmentation (i.e. too many regions are found) for sequences with high level of details (e.g. extern ambient). A colour based segmentation It is based on the assumption that a uniformly coloured region also moves uniformly, which is generally true for natural objects, but most objects consist of more than one uniformly coloured region or are textured. Because of this fact, this approach often leads to oversegmentation. For this reasons alternative aggregating criteria may be based on more complex features as texture signatures or dominant colour sets, thus exploiting diversity criteria to reduce the number of found regions. Another segmentation approach is based on edge detection and suffers from similar problems. Also, watershed segmentation is commonly used for the initial segmentation e.g. in [Tsai01]).

If the initial segmentation results in oversegmentation, an algorithm has to be used for merging regions with similar motion parameters, or adding parts of adjacent regions or single pixels to a region for which the motion has been determined, an approach called region growing. On the other hand, it may happen that the initial segmentation contains regions, in which more than one motion is present and

therefore suitable motion parameters cannot be found. The algorithm must then separate the region correctly, e.g. by clustering possible motion parameters [Duf95].

There are basically two kinds of methods for estimation the motion of a region of the previously found segmentation: indirect methods, that first estimate a dense motion vector field for a region and then try to fit a parametric model for the region to it, and direct methods, that estimate the parameters of the motion model directly from the image data [Tek95], [Duf96].

6.2.2.1 Indirect Methods

Indirect methods estimate a dense motion vector field employing a non-parametric model. Then, for each region of the segmentation, motion parameters are estimated from the motion vector field inside the region. The parameters of the motion model selected for the representation of a region's motion can, for example, be found by least square modelling. This method is very sensitive to outliers [Duf96], which is a severe problem, as the initial segmentation may produce regions containing more than one motion. Other approaches use clustering methods (as in [Wa94]) or a modified version of the Hough transform to cluster by motion parameters in Hough space [Nash97]. The parametric models can also be found using Bayesian methods.

Indirect methods suffer especially from weakly defined motion boundaries caused by smoothness constraints in the motion estimation and occlusion problems [Duf96].

6.2.2.2 Direct Methods

Direct methods estimate the parameters of the motion model, usually an affine (6 parameter) or projective linear (8 parameter) model, for each region found in the initial segmentation.

All the commonly used parametric methods can be used for parameter estimation (also block based methods) if the initial segmentation is block based.

A general problem for direct estimation of motion parameters is posed by small regions, which may not contain enough structure for reliably estimating a larger number of parameters. Such small regions may especially occur as a result of oversegmentation.

Many methods that employ direct segmentation based motion estimation use the estimated motion to refine the segmentation, and for the regions of the new segmentation motion is again estimated. This approach leads directly to joint motion estimation and segmentation, as described below.

6.2.3 Joint Motion Estimation and Segmentation

It makes sense to combine motion estimation and segmentation into one algorithm, so that both can mutually benefit from the other's result. The methods discussed before can be extended to joint estimation and segmentation techniques. This is done by sequentially repeating motion estimation and segmentation steps to refine segmentation and calculate motion estimates for the new regions, which are expected to be more suitable regions of support, as a whole region is subject to the same motion. Another class of methods formulates a Bayesian estimation criterion including both region and motion models. These methods are called simultaneous estimation and segmentation methods.

6.2.3.1 Sequential Estimation and Segmentation Methods

Basically, any combination of motion based segmentation and segmentation based motion estimation methods could be used for this purpose. However, some more specialized methods have been proposed and a few of them will be presented briefly in the following.

Global motion compensated Watershed segmentation and region merging

Tsaig and Averbuch [Tsai01] suggest to estimate the global motion in a first step and then compensate for it, so that moving regions and associated occlusion regions remain, as they are not only subject to the global motion. An initial segmentation is performed using watershed segmentation and regions are spatio-temporally merged to correct the oversegmentation. The method is based on the assumption that region adjacency does not change over time, which is formulated as a Bayesian criterion.

Luminance clustering and region merging and splitting

Dufaux et al. [Duf95] also suggest to estimate and eliminate global motion in a first step. The image is then prefiltered to have regions with constant intensity and sharp contours. The segmentation is performed by clustering on the luminance values and affine motion parameters are estimated for each region. Regions with similar motion parameters are merged, regions in which more than one motion is present are split by performing a static segmentation of the region and selecting subregions that produce high prediction errors, when the parameters estimated before are applied.

Oversegmentation and region merging

Dang et al. [Dang95] present an approach which is based on initial segmentation by intensity, expected to result in oversegmentation. Affine motion parameters are estimated for each region using a DPD criterion and regions with similar motion are merged. Additionally, the boundaries of the regions are adjusted after each region fusion.

Frequency domain approach

The approach presented by Krüger and Calway [Krü96] uses a block based segmentation and an affine motion model for each block. The affine motion parameters are calculated by a frequency domain method based on the shift property of the Fourier transform. If the motion in a block is not uniform, the block is split into four sub-blocks and so a quad-tree data structure for the representation of the segmentation is created.

Bayesian framework

Strehl and Aggarwal [Strehl00] propose a method using a Bayesian framework, which performs a parameter estimation and a segmentation step in each iteration. The labelling, i.e. the assignment of each pixel to a region, is done on the basis of a maximum a posteriori probability. For the estimation of a region's motion parameters, an affine model is used. The algorithm is initialised using either an initial segmentation or an initial set of motion parameters, which can be based on results from previous frames, domain specific knowledge or random guesses.

Expectation-maximization (EM)

The expectation-maximization algorithm [Moon96] can be applied to the motion segmentation problem, if the number of possible motion parameter sets is known or could be estimated [Wei97a]. The expectation (E) step of the algorithm, which is derived from the Bayes rule and is nothing else than the structural model of the Bayesian approach, assigns for each point a weight to each motion parameter set, based on how well the parameters fit the data. In the maximization (M) step, the motion parameters for each of the regions are calculated and each pixel influences the solution based on the weight estimated before.

6.2.3.2 Simultaneous Estimation and Segmentation Methods

Both methods presented in the following are based on a Bayesian estimation and segmentation criterion.

The method proposed by Bouthemy and François (e.g. discussed in [Duf96]) employs an energy function based on two a priori models. The region prior model expresses a priori assumptions about the segmentation, e.g. homogeneity of regions, and the motion prior model is an affine model, using a potential function that is an extension of the optical flow equation. The potential is not calculated for single pixels but for small blocks, as more information is required to estimate the parameters of the affine model. Convergence is assumed when the number of pixels changing regions in an iteration falls below a threshold. The segmentation is used as the initial state for the next frame to assure coherence over time.

Chang et al. (in [Cha97]), an earlier version of this approach is discussed in [Tek95]) use a structural model based on the displaced frame difference and also two a priori models, one for segmentation and one for motion. The motion model for each region has a parametric component (6 parameter affine or 8 parameter projective linear model) and an additional non-parametric (translational) component, called residual motion. The region model encourages formation of contiguous regions by punishing

outlying region labels. To initialise the algorithm, a dense motion vector field with a global smoothness constraint is calculated. The image is initially divided into small blocks and for each of them affine motion parameters are calculated based on the motion vector field. The Bayesian criterion is optimised using the highest confidence first (HCF) method [Li95].

7 Shot Boundary Detection

Many visual features change at shot boundaries. It is therefore crucial to detect shot boundaries before doing further analysis, i.e. to perform a temporal segmentation of the video. A shot is a continuous sequence of frames taken by a camera, delimited by shot boundaries. Shot boundaries are transitions between two shots. There are two basic types: abrupt transitions (cuts), which are the most common ones, and gradual transitions (fades, dissolves, wipes, etc.) [Kop01]. Cuts, fades and dissolves account for 99% of all transitions found in film and video material, so most work has concentrated on these types.

Although shot boundary detection has been an active research topic for many years, there exists no general purpose solution which works well on every type of video material and for every type of transition. Changes within a shot, such as fast motion or lighting changes often lead to false detection [Hanj02].

In the following, we will review the state of the art of shot boundary detection and discuss the most commonly used approaches and their relative strengths and weaknesses. Comparisons of different approaches and references to detailed descriptions of different algorithms can be found in [Lien01][Kop01][Bor96][VA03].

7.1 Shot Boundary Detection

Independently of the feature used, many of the methods discussed below require some kind of threshold to determine if a feature difference is caused by a transition or not. The definition of a suitable value for a threshold is a problem. Adapting thresholds to the content using the values of some feature (e.g. motion activity) within a time window detection accuracy, can be a solution.

Alternatives to the use of thresholds are statistical methods, which are performed on a whole image sequence and try to classify shots into frames containing transitions and those that do not. Clustering techniques are commonly used for this task. The drawbacks of these methods are that the data of a whole sequence has to be stored and processed at once, so that a decision about shot boundaries can only be done in retrospect. This means that all processing steps applied after shot boundary detection can only be done on the whole sequence.

Between the two classes discussed before, there are hybrid approaches, which may benefit from data stream clustering techniques, i.e. algorithms that operate on streams of data (such as realtime-extracted video features) utilising clustering methods requiring reduced amount of memory to store the intermediate structures needed to construct clusters during stream acquisition ([GMMMO03]).

The work in [Hanj04] identifies three inputs for deciding about the presence of shot boundaries: feature discontinuity (i.e. the visual difference between frames of the video), prior information (knowledge about the properties of the material, such as typical shot length) and discriminative information, which is modelled as a discriminative function, that relates feature properties to the occurrence of a certain type of shot boundary. The following overview is organised by the features used for determining discontinuity and discriminative information.

7.1.1 Colour/Intensity Based Approaches

A large number of approaches uses colour or a derived statistical measure (e.g. a histogram) to determine visual changes which are likely to be transitions.

7.1.1.1 Pixel Comparison

This is the simplest approach method which compares pixel values pair-wise. Some measure, e.g. the accumulated pixel difference over the image, is compared against a threshold to determine if a transition has been found.

This approach is sensitive to local changes caused by varying illumination and motion. Gradual transitions cannot be distinguished from motion with this approach.

Using second order derivatives for the of pixel or histogram differences has been proposed to make the detection more robust against illumination changes and object motion [Yuan04].

7.1.1.2 Block Comparison

Instead of comparing single pixel values, average values of blocks of pixels are compared. This makes the comparison more resistant against noise, slow local motion and local non-rigidity of objects. The work described in [Boch00] and [Cam98] is a shot detection technique based on this approach and on the use of an adaptive threshold to filter off the effect of local temporal activity and distinguish it from actual shot changes. Overlapping blocks can be used to reduce the problem of dealing with changes appearing at block boundaries.

7.1.1.3 Global Histogram Comparison

Histograms are a statistical measure commonly used to describe an image. For the purpose of shot boundary detection histograms are often calculated from colour spaces with less correlated components than RGB, such as the HSV colour space. The advantage of global histogram methods over pixel or block comparison methods is the stability against local motion and noise. Objects moving around in the scene will not change the global histogram.

Independently of the feature being used, the histogram will very likely change significantly at cuts. Gradual transitions will also cause gradual changes of the global histogram, a pattern that may be similar to histogram changes caused by motion, e.g. by a camera pan, where parts of the scenery disappear and others appear, also causing a gradual change of the global histogram. . Gradual transition detection problems become even harder when content before and after the transition is moving.

An approach to make histogram comparison more robust for gradual transition detection is to compare the histogram over a range of timescales [Heesch04].

The global properties of an image (such as mean and standard deviation) can be used to detect monochrome frames, which indicate fade ins/outs, if there is a gradual transition around the monochrome frame [Yuan04].

Many algorithms use a model of the global colour variations caused by flashlights in order to reduce the number of false detections in a postprocessing step.

7.1.1.4 Local Histogram Comparison

Local histogram methods have been introduced to handle cases where local changes severely influence the global histogram, e.g. newly appearing objects or text inserts in a part of a screen. Local histograms are calculated (typically over blocks) and a transition is assumed only if a large number of the local histograms supports it. Smaller blocks increase the resistance against local changes, but are more sensitive to motion. Some approaches apply weighting to the different local histograms in order to be more robust against object motion, which often appears in the centre of the image [Volk04].

7.1.2 Edge Feature Based Approaches

Edge features have some advantages over colour/intensity features. They are insensitive to moderate global illumination changes and less sensitive to local illumination changes. In the case of rigid object motion, the edge structures are not changed but just displaced and their statistical properties remain the same. Abrupt transitions cause global abrupt changes of the edge features, while transitions and fades cause patterns of decreasing and increasing edge energy. These patterns can be compared to models of certain types of transitions (cf. [Yu97]).

Edge features alone are not sufficient, as shots showing similar content (e.g. persons) will have similar edge features. It is necessary to use colour/intensity based methods additionally.

7.1.3 Motion Based Approaches

The concept of optical flow is based on the assumption that pixel intensities do not change along motion trajectories. After estimating the optical flow, the image intensities are compared. Abrupt or gradual changes of a large number of pixel intensities indicate a transition.

The main drawback of this method is that the optical flow estimation is unreliable in the presence of illumination changes and motion phenomena like occlusion and is likely to fail during gradual transitions. The same is true for methods using block based motion estimation [Lien01].

Compensation of the global motion in the image sequence is used to improve the performance of pixel intensity or histogram comparison methods (cf. [Que04]).

7.1.4 Feature Tracking Based Approaches

To exploit the information contained in object motion to discriminate between shot boundaries and scenes with high motion activity without the computational cost and maybe instabilities of motion estimation, feature tracking based approaches have been proposed. Feature tracking approaches are based on the consideration that both abrupt and gradual changes will cause disappearance of some existing features and appearance of new features.

The method proposed in [Whit04] is restricted to the detection of hard cuts. It seems, however, that this class of approaches has potential for the detection of gradual transitions, as they are even more difficult to discriminate from scenes with high visual activity.

7.1.5 MPEG Compression Domain Approaches

Performing shot boundary detection in the MPEG compressed domain has two main advantages: It is not necessary to decompress the video before processing and no or fewer new features have to be extracted. The main disadvantage is of course, that only MPEG encoded video data can be processed. However, this disadvantage is becoming less important due to the increasing processing power of computers, so that an ad-hoc MPEG encoding with the only purpose of extracting these features could easily be regarded as feasible with no sensible extra requirement of hardware capabilities in many practical cases.

Rather, the efficiency of such an approach may heavily depend on other factors as the source images quality (e.g. presence of coding artefacts, like tiling effect, introduced by some early digitisation of content) and specific criticality of images with respect to MPEG encoding (e.g. high temporal and spatial activity) which tend to introduce image distortions that might consistently affect the measurement. Besides, starting from already encoded sequences may suffer from the choice of the particular GOP (Group of Pictures) structure at encoding time for potentially different purposes (e.g. archiving).

In the MPEG domain, there are both colour and motion based approaches, using DC values, DCT coefficients and motion vectors respectively. The colour based approaches in the MPEG domain are basically block comparison methods, the use of DCT coefficients is similar to using local histograms calculated over (in this case very small) blocks. The motion vectors available in the MPEG stream are usually generated for the purpose of coding efficiency and not for accurate description of true motion.

The approach proposed in [Pet04] has specialised detection strategies for hard cuts, dissolves and wipes. For the first, the commonly used differences in the DC images as well as the edge energy are used, for the latter the lines created by wipes are detected using a Hough transform.

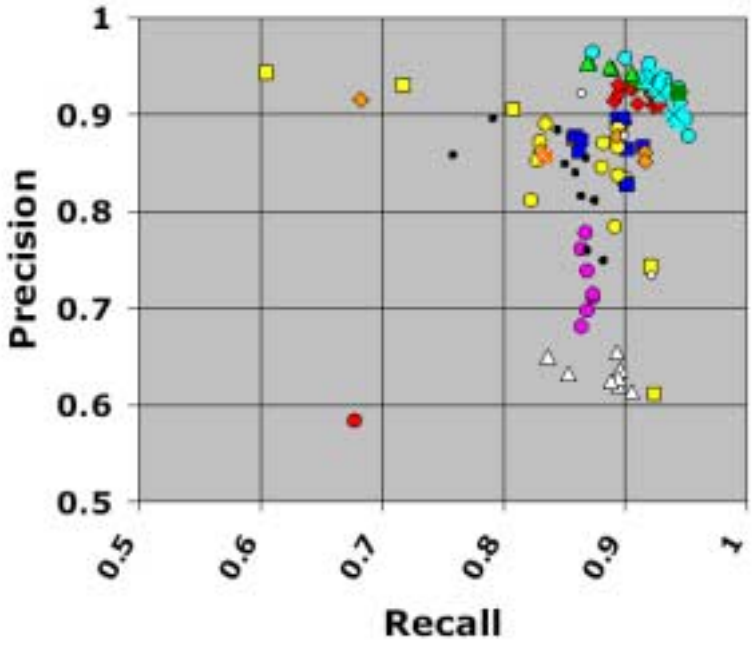
7.2 Performance of Shot Boundary Detection Approaches

The TRECVID Video Retrieval Evaluation [Trecvid] aims at objectively comparing automatic content-analysis and retrieval systems. One of the oldest tasks is shot boundary detection on a defined video set. The results are compared to a human annotated ground truth.

Figure 3 shows the result of the TRECVID 2004. The figures show the precision and recall measures for the different algorithms (different symbols indicate different algorithms, the different results from one algorithm are due to different parameters). Precision measures the fraction number of correctly detected transitions among all the detected transitions, recall measures the fraction of detected transitions among all actual transitions in the reference.

It can be seen that the algorithms of abrupt transitions (cuts) are quite elaborated and provide satisfying results. In contrast to that, there is still room for improvement in the detection of gradual transitions, as no algorithm exceeds recall values of 0.85, and only those with lower recall values reach precision scores beyond 0.85 (i.e. they only report gradual transitions when the likelihood is very high).

In TRECVID 2004, information about the runtime performance of the algorithms has been collected for the first time. The fastest algorithms working in the uncompressed domain are up to 2-3 times faster than real time, those working in the compressed domain up to 20-30 times real time.



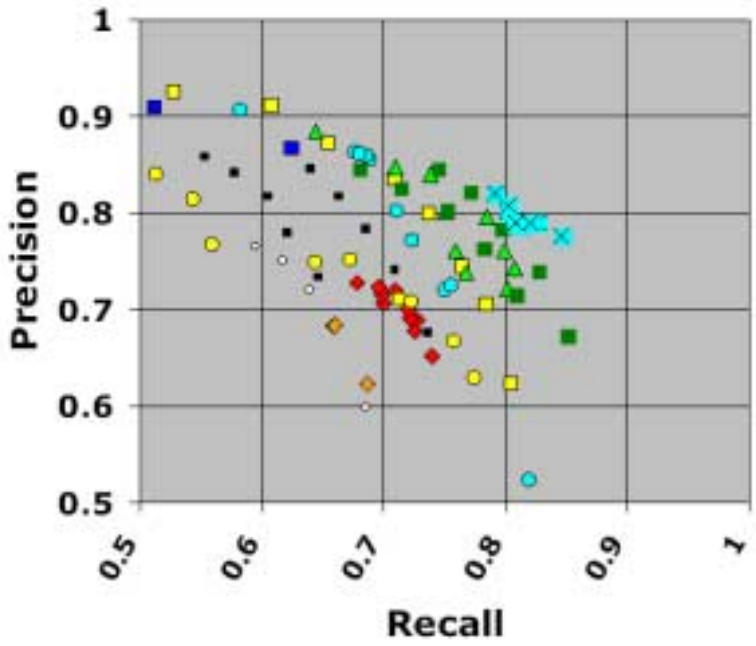


Figure 3: Precision and recall performance of shot boundary detection approaches. The upper figure shows the results for abrupt transitions (cuts), the lower figure for gradual transitions (from [SOTV04]) Different symbols indicate different algorithms.

8 Video OCR

Compared to normal OCR, video OCR is a quite complex task because there are usually a lower resolution, complex backgrounds and the text may appear in different slant, tilt and lightning condition. Video OCR is very often specialised to a certain domain where domain knowledge can be used. For example, news broadcasts where mainly overlay text is used for OCR. A general text detection in video is more complex, because text may appear for example on road signs, shops (shop name), street names and so on.

A generic video OCR tool consists mainly of three steps:

1. Text detection
2. Text segmentation and enhancement
3. Classical OCR

The task of text detection is to find text occurrences in images and videos. In the text segmentation step the image areas containing text are prepared for OCR. And in the OCR step standard OCR software can be used to recognise the text.

There has been a quite recent survey from Lienhart [Lienhart03] on the subject of video OCR which gives a more detailed insight into this subject.

A good overview on papers can also be found in [TextSeg].

8.1 Text Detection

The task of text detection is to detect occurrences of text in images and videos. Text has some unique features which may be utilised. These features are mainly exploited by statistical analysis.

The text detection approaches can be classified into three categories.

1) Connected component-based methods
2) Edge-based methods

3) Texture based methods

The connected component based methods can locate text quickly but have difficulties when text is embedded in complex backgrounds or touches other textured areas.

The edge based methods generally decompose text regions by analysing the projection profiles of the edge intensity maps. This kind of methods can hardly handle large-size text.

In this report we deal mainly with the category of texture based methods where the textual features are used.

The most common feature based text detection algorithms are working on a fixed scale and fixed position. To overcome this limitation sliding widows and multi resolution images are used.

Clark and Mirmehdi [Clark00] proposed a method based on statistical properties of local image neighbourhoods for the location of text in real-scene images. There text detection is to a large degree invariant to orientation, scale, and colour.

In video text detection the temporal redundancy is used to increase the text localisation. For example the background may change from frame to frame. Also false text detection can be found by looking on multiple frames.

Text tracking is used to localise the same text over a sequence of images.

The last step in text detection is a post processing step. For example, text must be visible for at least two seconds to be meaningful to a human viewer. In the post processing step the text appearances are reduced to the areas which are highly likely to contain text.

For a detailed comparison of text detection algorithms see [Lienhart03].

8.2 Text Segmentation

The task of text segmentation task is to prepare the areas of localised text for OCR by means of quality enhancement of the image.

One problem in text recognition in video is the resolution. In order to achieve good text recognition results, characters should have at least a size of 40 pixels, while a size of 11 pixels is quite common in video. To overcome this problem a sub-pixel accurate re-scaling of the original text bitmaps to a fixed target height is used.

In the video domain the same text may appear over multiple frames. This allows dealing better with complex backgrounds and lightning conditions and to use techniques like super-resolution from motion for enhancement.

8.3 OCR

After text detection and segmentation the areas containing the text are provided to “ordinary” standard OCR software. The standard OCR packages have reached a quite good level of maturity. OCR itself is not be covered within this report.

9 Face Detection and Recognition

Face detection is an important step in almost any face recognition system. The task in face detection is to find all regions in arbitrary sized images containing human faces. The detection is still a challenging task since the faces may appear in different scales, orientations, rotations, and head poses. Different illumination and complex backgrounds are further challenges. Last but not least, faces are non-rigid objects and a lot of variations due to varying facial expressions are possible.

In the past years face detection has been a challenge for many researchers and a lot of face detection methods have been proposed in the literature. There have been some surveys [Yang02], [Hjelmas01] on this topic and some of them are quite recent. Those surveys are the basis of this report. This report should act as an aggregation of those surveys and more recent work which is not covered within these surveys will be discussed as well.

A good starting point for face detection information is the face detection homepage from Robert Frischholz [Frisch].

9.1 Face Detection Approaches

There are different approaches in face detection. The methods can roughly be categorised in some groups however many methods can be categorised in several ways.

- Knowledge-based methods (top down approach)
- Feature-based methods
- Feature invariant approaches
- Template matching methods

9.1.1 Knowledge-Based Methods

The key in the knowledge-based methods is that the human knowledge of how a typical face is constituted is encoded in some way. Rules encode this knowledge. Typically those rules model the relation between facial features. This is a top down approach in face detection.

Pros:

- It is easy to find simple rules to describe the features of a face and their relationships
- Based on this rules the facial features in an image are extracted first and are used to identify face candidates

Cons:

- It is difficult to translate human knowledge in rules
- It is difficult to find faces in different poses
- It is difficult to detect multiple people or people in complex background

9.1.2 Feature-Based Methods

This is a bottom-up approach in the sense that facial features (eyes, nose, mouth, nostrils, etc.) are detected first. The features are detected using some low level feature analysis.

Low level features:

- Edges
- Grey-levels
- Colour (skin colour)
- Shape
- Texture
- Intensity

Skin colour is a good key to find face candidates. Skin colour across ethnic groups are not a problem since the human skin colour falls into a small range in different colour spaces regardless of races.

Those low level features are analysed and grouped to form face candidates. Those candidates are verified by further analysis.

Leung et al. [Leung95] proposed the use of random label graph matching to group the features. An other feature-based system is used by Yow and Cipolla [Yow96]. Here the features are detected using elongated Gaussian derivative filters. This filter is able to detect features like corner of the eyes, mouth and nose, even in the case when faces are not frontal.

Pros:

- The features are invariant to pose and orientation

Cons:

- It is difficult to locate facial features due to corruption (illumination, noise, occlusion)
- Many features are also detected in the background

9.1.3 Template-Based Methods

Template based methods are using standard face patterns of a fixed size and this acts as a search window which is scanned over the image. At each location the corresponding part of the image is classified as a face or non-face pattern. To detect faces of different scales, either the input image is scaled down or the size of the template is adjusted. Many standard pattern recognition methods are applied in this approach. The templates are predefined based on edges or regions or deformable templates based on facial contours (snakes) are used. Note that the templates are hand coded and not learned.

Pros:

- Simple to implement

Cons:

- Templates needs to be initialised near the face image
- It is difficult to enumerate templates for different poses
- It is difficult to deal with different scales

9.1.4 Appearance-Based Methods

Other than in the template based methods where the templates are predefined the “templates” in appearance based methods are learned from examples in images. In general the appearance based methods relay on techniques from statistical analysis and machine learning to find the relevant characteristics of faces and non-faces.

In the analysis many different approaches for the classifiers are used

- Neuronal networks
- Principal Component Analysis (PCA)
- Support vector machine (SVM)
- Distribution based method
- Naive Bayes classifier
- Hidden Markov model
- Sparse network of winnows (SNoW)
- Kullback relative information
- Inductive learning
- Adaboost
- …

Pros:

- Uses powerful machine learning algorithms
- Has demonstrated good empirical results
- Fast and fairly robust
- Extended to detect faces in different pose and orientation

Cons:

- Usually needs to search over space and scale
- Needs a lot of positive and negative examples
- Limited view-based approach

9.1.5 Video-Based Detector

Face detection in video uses the motion information. This can be done by frame differencing and background model subtraction. This can reduce the search space dramatically.

Pros:

- Detecting faces in videos is easier than in still images
- Cues like motion, depth, voice can be used to reduce search space

Cons:

- Need to efficient and effective methods to process multiple cues
- Data fusion

9.1.6 Face Detection Software

Free Face Detection Algorithms & SDKs [Frisch]

BuFaLo Face Localisation OCX	http://www.geocities.com/fritzfra2001 Biometric base Unit for Face Localisation OCX implements the work of Viola & Jones, "Robust real-time Object Detection". Including the enhancements by Lienhart et al. Uses the OpenCV library.
Face Tracking DLL from Carnegie Mellon	http://amp.ece.cmu.edu/projects/FaceTracking Face tracking using colour matching combined with deformable templates.
Real-time face detection program from FhG-IIS	http://www.iis.fraunhofer.de/bv/biometrie/download/index.html Demo from the Fraunhofer Institute IIS, Germany. Shows face tracking and detection using edge orientation matching. Fast multi-face finding capabilities. Executable only.
OpenCV	http://www.intel.com/research/mrl/research/opencv Intel initiates an open source community for computer vision. C++ source code for face recognition, motion tracking, and many others.
Computer Vision Source Code	http://www-2.cs.cmu.edu/afs/cs/project/cil/ftp/html/v-source.html Useful collection of image processing code.

Commercial Face Detection Applications & SDKs

Most commercial face detection SDKs offer face recognition as well. For commercial tools see chapter 9.2.4 for available products.

9.2 Face Recognition

In the past several years face recognition has received special attention. The reasons for it are that technically the face recognition has reached a certain level of quality and on the other hand the law enforcement agencies have shown more interest in it. Face recognition is still fare away from being solved completely especially in outdoor environments.

Face recognition in general is the task to identify persons in still images or video using a face database holding the persons which can be identified. Face detection is used to segment the image into face and non face regions. In the face regions facial features are extracted followed by recognition and verification steps. The face feature extraction may also be part of the face detection and could be shared between detection and recognition.

There has been a recent face recognition survey [Zhao03]. This report is based on that survey.

9.2.1 Face Recognition in Still Images

There are many different methods to recognise faces from still images. This chapter should provide a short overview over current methods.

9.2.1.1 Recognition from Intensity Images

These methods can be categorised in classes.

- Holistic matching methods
  These methods use the whole face as input to the recognition system.
- Feature-based matching methods
  These methods are using local features such as eyes, nose, and mouth. Their locations and local statistics are further processed.
- Hybrid methods
  Both holistic and feature-based approaches are combined.

Holistic Approaches

- Principal-Component Analysis (PCA)
  - Eignefaces
  - Probabilistic eigenfaces
  - Fisherfaces/subspace LDA
  - SVM
  - Evolution pursuit
  - Feature lines
  - ICA
- Other representations
  - LDA/FLD
  - PDBNN

Feature-based methods

- Pure geometry methods
- Dynamic link architecture
- Hidden Markov model
- Convolution Neural Network

Hybrid methods

- Modular eigenfaces
- Hybrid LFA
- Shape-normalised

- Component-based

9.2.2 Video-Based Face Recognition

In video based face recognition the main techniques are face segmentation and pose estimation, face tracking, and face modelling.

9.2.2.1 Basic Techniques of Video-Based Face Recognition

Face Segmentation and Pose Estimation

The first step is to segment moving objects from an image sequence. Colour may be used to speed up the process of finding face regions. After candidate regions are found, still image face detection techniques can be applied to locate faces. From the face regions facial features can be extracted which are used for pose estimation. With this information a virtual frontal view can be computed.

Face and Feature Tracking

After a face has been detected, the faces and their features are tracked. The facial image over the time is used to reconstruct a face model. The tracked features are used for facial expression recognition and gaze recognition.

Face Modelling

This includes 3D shape and texture modelling.

9.2.3 Video-Based Face Recognition

Early versions of video face recognition used still-image based techniques on the detected faces in the video. An improvement was using tracking and creating virtual frontal views. Also voting based on the recognition results from each frame is used. The next generation of face recognition systems used multimodal cues. For example body motion characteristics and voice. More recently spatiotemporal methods are used.

9.2.4 Face Recognition Software

Face recognition Algorithms [EvalFaceR]

This web site is a resource for researchers developing face recognition algorithms. It provides a standard set of well known algorithms and established experimental protocols. The goal is to provide a firm statistical basis for drawing conclusions about the relative performance of different algorithms and to better explain why algorithms behave as they do.

Commercial Face recognition Products

FaceIt from Visionics	http://www.faceit.com They have a FaceIt SDK to develop face recognition applications. FaceIt DB is can be used to search in large databases.
Viisage Technology	http://www.viisage.com They have the Viisage FaceTOOLS SDK
FaceVACS from Plettac	http://www.plettac-electronics.com Mainly for access control solutions.
FaceKey Corp.	http://www.facekey.com Mainly for access control solutions.
Cognitec Systems	http://www.cognitec-systems.de They have FaceVACS-SDK to develop applications and have also access and border control applications.
Keyware Technologies	http://www.keyware.com
Passfaces from ID-arts	http://www.realuser.com
ImageWare Software	http://www.iwsinc.com They have IWS EPI Builder which is a software developer's kit (SDK)
neven vision	http://www.nevenvision.com They have SDK's for face recognition and facial feature tracking.
BioID sensor fusion	http://www.bioid.com They offer a SDK
Visionsphere Technologies	http://www.visionspheretech.com
Biometric Systems, Inc.	http://www.biometrica.com
FaceSnap Recorder	http://www.facesnap.de
SpotIt for face composite	http://spotit.itc.it
Imagis Technologies	http://www.imagistechnologies.com They have a SDK
identix	http://www.identix.com
NEUROtechnologija	http://www.neurotechnologija.com They have the VeriLook SDK for face recognition.
A4Vision	http://www.a4vision.com/ Vision Access Software Developer Kit (SDK) The Vision Access Software Development Kit (VA SDK) enables application development for configuration, network monitoring, deployment and integration of A4's security solutions. Using the VA SDK you can quickly and easily integrate into existing physical access and network security solutions. The VA SDK can be used to seamlessly integrate into a wide-range of third party applications as well as new application development.

10 Defect and Quality Analysis

This section contains a survey of the state of the art of visual defect and quality analysis algorithms and tools.

Defect and quality analysis tools can be used to automatically extract metadata from the essence for the following purposes:

- Indexing defect and quality descriptions for quality assessment in media archive applications. The defect and quality metadata is then searchable by content-based methods.
- Describe defect and quality over media time to support the process of restoration or the estimation of restoration effort.

Defects can be grouped by the modality (visual/audio) and the type of media (e.g. film or video). Visual defects can be grouped as follows:

- Film defects: Examples of film defects are flicker, unsteadiness, scratch, dust, dirt, grain, missing frames, mould, dye fading, vinegar syndrome and dirty splices.
- Video defects: Examples of video defects are various kinds of drop-out, jitter, ghosting, overshoot, head-clogging and cross-chroma. An overview of defects related to video and video equipment related can be found in [Brava].
- Media independent defects: Additionally there exists a group of defect and quality measures which are applicable across media, e.g. noise and blur. Finally there is a group of quality/defect measures related to effects caused by encoding/decoding of audiovisual media, e.g. tiling or edge busyness.

Apart from the origin of certain defects an important criteria for measuring defects is the measurement paradigm. There are two main paradigms, and some shades in between. One defect/quality measurement paradigm relies on ground truth data (also called reference based measurement), the other measurement paradigm does not use any reference data (also called non-reference based measurement).

10.1 Reference based quality/defect analysis

In the area of video broadcast/delivery there are several activities which aim at finding quality/defect measures for transmission systems. Some of them are already standardised, others are being standardised. Such activities are driven by expert groups (e.g. VQEG) which are working closely together with standardisation organisations (e.g. ITU, ANSI).

10.1.1 VQEG

The main purpose of the Video Quality Experts Group [VQEG] is to provide input to the relevant standardisation bodies responsible for producing international recommendations regarding the definition of an objective Video Quality Metric (VQM) in the digital domain. The group was formed in October 1997. The majority of participants are active in the International Telecommunication Union (ITU) and VQEG combines the expertise and resources found in several ITU Study Groups.

The VQEG finished in 2000 a first test phase (project FRTV Phase I) where different objective video broadcast quality metrics have been compared. Statistics between the objectively measured quality decrease (between an original essence and an encoded, transmitted and decoded essence) and the subjectively perceived quality decrease have been done. Reports on the results of the phase I can be found in [FRTV1a] and [FRTV1b]. The phase 1 report concludes that no one of the proposed objective metric reflects the subjective test results properly and thus cannot be proposed for standardisation.

In 2003 a second test phase (project FRTV Phase II) has been finished. With a similar test procedure but different objective metrics VQEG concluded that some objective quality metrics in this test perform well enough to be included in normative sections of recommendations. The results can be found in [FRTV2].

10.1.2 ANSI T1.801.03-1996

The ANSI standard T1.801.03-1996 [ANSI801-96] is dealing with how to measure video quality of broadcast delivery and video conferencing systems. The standard proposes measures for certain degradations introduced by transmission systems. Measurement is done by comparing transmission input and output. Some measures from this standard were used during the PRESTO project for evaluating film and video encoding and decoding quality, a detailed report can be found in [Schall02].

The standard is available since 1996 (ANSI T1.801.03-1996), with a revision published in 2003 (ANSI T1.801.03-2003). It was included in two International Telecommunication Union (ITU) recommendations in 2004 (ITU-T J.144R and ITU-R BT.1683). The measurement paradigm is known as "reduced-reference" video quality measurements (cf. [ITU-T_J.143][ITU-T_J.143]).

Different quality measures are standardised, among others there are measures for blurring, block distortion (tiling), error blocks, added noise, added edge business and jerkiness.

The VQEG metrics as well as the ANSI T1.801.03 measurements are based on the availability of undistorted reference essence and distorted encoded/decoded essence. Although the availability of both is usually not the case for archived essence, algorithms and measures developed by these standards are an interesting base for research and development of non-reference based quality/defect measures.

10.2 Non-Reference based quality/defect analysis

Non-reference based analysis tries to find measures audiovisual quality/defects without using of any ground truth data, e.g. no undistorted reference video to compare against is necessary. Because of this independence on ground truth data this approach has a wider application area than the reference based approach. On the other hand, the unavailability of ground truth data makes the calculation of quality/defect measures more difficult than in the reference based approach.

It is often the goal to define defect/quality measures that correlate with human perception. Typically this requires expensive subjective tests with following statistics calculation about correlation between subjective perception tests and objective measurements.

In [Cou01] an objective measure that predicts the visibility of the well-known blocking effect in discrete cosine transform (DCT) coded image sequences is presented. The proposed measure is based on a visual model accounting for both the spatial and temporal properties of the human visual system. The input of the visual model is the distorted sequence only. Psycho-visual experiments have been carried out to determine the eye sensitivity to blocking artefacts, by varying a number of visually significant parameters: background level, spatial, and temporal activities in the surrounding image.

In [Far05] perceptual analysis of video impairments is done for the case that blocky, blurry, noisy, and ringing synthetic defects are combined. The influence of combination of defects onto human perception is studied. The authors found that if a video is suffering from both noise and blocking defects, blocking is perceived less strong than in the case where no noise is present in the video. This means blocking that is perceptually decreased by noise. If a video is suffering from blurring and from blocking defects, blocking is perceived stronger than in the case where the video is not blurred, i.e. blocking is perceptually increased by blur.

Methods for non-reference based video quality assessment are proposed in [Wang00b], [Bovik00], [Gast01], [Knee01], [Wu97] and [Cav00]. They focus on objective measurement of blocking and other DCT defects.

[Bra96] describes a spatiotemporal model of the human visual system (HVS) for video imaging applications, predicting the response of the neurons of the primary visual cortex. The model can be used as the basis for building quality metrics, e.g. a quality metric for coded video sequences is presented.

[Xin02] proposes a blind image quality assessment method to appraise the image quality by three objective measures: edge sharpness level, random noise level and structural noise level. They jointly provide a heuristic approach of characterising the most important aspects of visual quality. Various mathematical tools are investigated (analytical, statistical and PDE-based) for accurately and robustly estimating those three levels.

[Cav02] describes a content independent, non-reference based sharpness metric. In this approach an edge profile is created by detecting edge pixels and enclosing them with 8 times 8 pixel blocks. For

each block the sharpness is computed using the kurtosis of the DCT. The final metric is the average sharpness of the blocks in the edge profile. A combination of spatial and frequency domain information is exploited.

In [Cha05] some measures for the colour distribution in moving images are given with the aim of characterising dye fading.

In the area of defects originating from film and video medium, research has focussed until now on spatiotemporally very detailed analysis for the purpose of digital restoration, e.g. detailed dust, flicker, grain, unsteadiness, scratch or video dropout detection. Among others, such restoration approaches can be found in [Bess04], [Buis03], [Joy00], [Kok98], [Schall99] and [Vla04]. This research is typically not targeted towards large scale quality and defect analysis for digital archives, which generally requires faster algorithms or implementations.

Part B: Content Analysis Tools for Audio and Speech

Before introducing content analysis tools for audio and speech, is useful to remember that an important part of signal processing methods and approaches available in literature are common to many different domains. Audio and visual domain share a large part of this background, and according to the introduction for low-level visual features illustrated in part A (see Section 5.1), the accuracy and robustness of audio content analysis tools basically depends on the underlying low-level feature extraction. The formal approach and the requirements for the descriptors created from the extracted audio features are the same, while representation, extraction and comparison methods between feature descriptors remain the main way to feed different high-level algorithms.

11 Low-level Audio Features

Depending on the specific analysis tool, different subset of low-level features may be considered. Some features may be redundant or useless, conversely other features may be mandatory in order to obtain an effective descriptor.

An important issue that deeply impacts into feature extraction is the size of the analysis window (both in temporal domain and in frequency domain). Normally the audio stream is temporally split in small regions (or segments) on which feature algorithms are applied. The size of this window is often a critical choice [Davy02] in order to obtain a meaningful feature.

Depending on the analysis tool, the typical size of an analysis window varies from 1 second for a macro-segment to small sub-segments of 10 milliseconds.

The following paragraphs illustrate commonly used audio features.

11.1 Short Time Energy (STE)

It is a time domain feature and is defined as the energy of the signal inside the analysis window. It provides a convenient representation of the signal amplitude over time [Zhang98a].

The prefix “Short Time” refers to the dimension of analysis window compared to other standard energy measurements

Usage: silence detection, speech/music classification and distinguishing voiced/unvoiced speech components.

11.2 Low Short Time Energy Ratio (LSTER)

This derivative time domain feature from STE [Lu02] is based on average STE values taking benefits from multiple window measurement and highlighting frames with low (i.e. lower than average) STE.

Usage: silence detection, speech/music classification and distinguishing voiced/unvoiced speech components.

11.3 Zero Crossing Rate (ZCR)

It is a time domain feature and represents a simple measure of the frequency content of the signal. It is obtained by counting the number of sign changes of the signal (“zero crossing”) inside the analysis window.

Usage: speech/music classification and distinguishing voiced/unvoiced speech components.

11.4 High Zero Crossing Rate Ratio (HZCRR)

This derivative time domain feature from ZCR [Lu02] is based on average ZCR values taking benefits from multiple window measurement and highlighting frames with high (then average) ZCR.

Usage: speech/music discrimination

11.5 Spectral Flux

It is a frequency domain feature obtained by measuring the average spectrum variation value between adjacent analysis segments.

Usage: speech/non speech and music/environment classification

11.6 Band Periodicity

It is a frequency domain feature obtained by a subband correlation analysis [Lu02]. For each selected subband the maximum local peak of the normalised correlation function is extracted.

Usage: music/environment discrimination

11.7 Median Frequency

It is a frequency domain feature (a.k.a. centroid frequency) the extract a kind of centre of gravity for the spectrum obtained with a typical frequency transformation (like DFT).

Usage: wide-band/narrow-band signal classification and speech gender (male/female) detection

11.8 Mel frequency Cepstral Coefficients (MFCC)

It is a frequency domain feature where the logarithm of the power spectrum is computed (using DFT). Then logarithmic spectral coefficients are perceptually weighted by a non linear frequency scale map. The final stage is a further DCT transform obtaining the cepstral coefficient [Molau01].

Usage: extracting formants from voiced phonemes in speech recognition

11.9 Fundamental frequency

It is a frequency domain feature (a.k.a. pitch detection) obtained by detection and extraction of the fundamental peak from the frequency spectrum. This feature is correlated to the presence of harmonic components in instrumental sounds or in voiced components of speech.

Usage: voice/unvoiced speech

12 Use of Low-Level Audio Features

The extracted low-level audio features can create different descriptors that can be used either as an input prerequisite for the extraction of higher level information or they can be used directly for similarity matching query. Another use is the use of low-level feature descriptors directly for media indexing purposes.

12.1 Extraction of higher level information

A general categorisation be presented as:

- Segmentation/classification methods
- Clustering methods
- Pattern retrieval methods

Two different approaches for segmentation, detection and classification methods are presented in literature: parametric methods assume that input signal follow some statistical/mathematical model, conversely non-parametric models make no assumption on the input signal.

Moreover, each algorithm requires at least a moment where a decision is taken. These decision schemes can be referred as parametric scheme (using Bayesian theory) where the optimal decision scheme in the space of the unknown model parameters is calculated from the knowledge of statistical distribution of the model parameters, or non-parametric schemes, consisting of finding the best similarity measure that best achieves correct classification. Non-parametric schemes require a training phase for design the best similarity measure. This approach is referred to as Learning Theory in literature.

12.2 Segmentation/Classification methods

The segmentation/classification problem has be handled in different ways that can be summarised in:

- Rule based methods: A typical approach used by segmentation algorithms takes advantages from thresholding techniques applied directly to the feature descriptors. The input audio stream is segmented in homogeneous regions where the descriptor manifest a steady state (or at least doesn't manifest a rapid change of feature values). These algorithms don't take care of the content of the segmented region. This approach allow a fast processing (real-time processing is simply achievable) but relies on the choice of the threshold. In order to improve the performance and reduce threshold sensitivity adaptive thresholding techniques are used, which obtain very high scores (more then 95%), when segmenting with silence/non-silence or speech/music paradigms.
- Metric based methods: segmentation and classification is achieved by evaluating the distance from different stream segment, where distance refers to one of many different methods such as Kullback Leibler distance [Kemp00], likelihood ratio [Gish91], entropy loss [Kemp00], and Bayesian Information Criterion [Schwarz78]. These approaches don't need prior knowledge on audio classes and can thus be applied to real time segmentation.
- Explicit model methods: models are built from a given set of acoustic classes (like silence/music/noise/male speech), then the audio stream is classified. The common part for classification algorithms is the internal classification model based on classes. Each class can simply describe a single "content-type" (as silence/music/speech/noise rough classification) or can contain and aggregate multiple subclasses (like female/male or narrow-band/wide-band for speech class) obtaining a hierarchical structure. Different approaches such as GMM (Gaussian Mixture Model) or HMM (Hidden Markov Model) can be used for class description and modelling while the implementation of the classifier can rely on k-NN (k-nearest-neighbour) algorithms, maximum likelihood principle or Viterbi Algorithms or on multi-stage approaches (a threshold rule-based algorithm is applied after a pre-classifier stage [Lu02]).

12.3 Pattern retrieval methods

Pattern retrieval methods represent a different use of segmentation/classification algorithms. These methods often follow the "query by example" paradigm, where many sets (at least one) of relevant audio patterns must be supplied allowing an initial training phase of the Gaussian Mixture Models (GMMs). A valuable importance must be given to the pattern modelling technique used since it is possible to build models that can take (or not) into account also the temporal course of the samples. The results can be a pure statistical model or temporal model using HMM (Hidden Markov Model) leading to difference matching criteria: the temporal model leads to a "trajectory matching" inside the feature space while the statistical-only model leads to a similarity matching.

A relevant use of pattern retrieval methods can be named as "sound spotting", where the set of relevant patterns composed by interesting items like jingles, program intro or endings, is use by a real time algorithm that supplies a specific-use segmentation like separating news program from weather report or advertising.

12.4 Clustering Methods

These methods refer to the important task of grouping together different segments (produced by classification or segmentation methods) across the analysed stream. The cluster creation criteria is generally based on similarity measures, but it is related to the specific classification (or segmentation) domain: a rough acoustical classification (like silence/music/speech/noise) allows different clustering approaches for segments classified as music (e.g. on specific instrument presence), while speech segments clustering takes advantage from speaker modelling techniques. Although different clustering techniques takes advantage from very different modelling tools, clustering methods are normally based on one of the following approaches:

- Bottom-up approach: segments are merged following divergence criteria using a distance measure (like k-NN or ML (Maximum Likelihood))
- Top-down (hierarchical) approach: segments are assigned inside a defined node hierarchy using a distance measure (direct maximisation of MLLR in [Johson99]) or using an adaptive hierarchy construction [Zhang98b].

13 Automatic Speech Recognition (ASR)

The finest level of feature extraction from audio stream is represented by automatic speech recognition. The abbreviation ASR (Automatic Speech Recognition) refers to multiple cross-knowledge and application domains (like acoustic, phonetics, linguistic and lexical domains) where many different tools are used jointly forming a complex infrastructure. The purpose of this chapter is to give a brief and general overview on the building blocks of an ASR system and to present realistic computational costs for each of the key steps in he process.

13.1 Audio Segmentation/Classification

A requirement of an ASR system is a correct input classification in speech/non-speech segments. These classification allows the rejection of non-speech segments and thus saves computation costs. Further classification of speech segments is based on signal bandwidth (narrow-band/wide-band).

Computational cost: very low – can be performed faster than real time (10x faster than real-time)

13.2 Speaker Segmentation/Clustering

Speech segments are segmented according to the specific low-level features discriminating gender (male/female speaker) and allowing an accurate modelling. MFCC (Mel-Frequency Cepstral Coefficients) allow a statistical modelling using Gaussian Mixture Models (GMM) and a temporal modelling can be achieved by using Hidden Markov Model (HMM) (in [Lu02] a Linear Spectral Pairs measure is used for precise speaker boundary detection). Gaussian Mixture Models (GMM) provide also an accurate method for speaker identification.

Computational cost: very low – can be performed faster than real time (10x faster than real-time)

13.3 Speaker Identification

The speaker identification task requires a database of speaker models. These speaker models can be created during a training phase where some material (few minutes) for each speaker must be supplied.

Major important drawbacks for generic identification systems (like speaker identification) are:

- unknown objects are always mapped within the finite object space, thus generating a critical identification noise. In other terms is not possible to have an “unknown speaker” until is not possible to identify an “unknown speaker” useful for model training
- computational costs grow not linearly with the speaker cardinality space

- maintenance of speaker database is required to avoid a slight but constant model growth
- Computational cost: depends on speaker database – can be performed in real-time (with 20-50 speaker set)

13.4 Speech Transcription

During speech transcription each speaker segment (turn) is split in many sub segments (sentences) according to pauses or other speech topics. During this phase the following models are used:

- Lexical Model: The lexical model consists of a collection of words (typical size for a lexical model could be 65000 words) where each word is mapped with its phonetic representation.
- Language Model: The language model is normally obtained from a training phase using a huge corpus of text from the application domain for the ASR (e.g. broadcast news, forensic reports). From the selected corpus the joint probability distributions for bi-grams and tri-grams (word sequence composed by two or three words) are extracted. The language model supplies a statistical relationship between words used in the application domain. An important reduction of word error rate (up to 10%) is achieved by matching the phonemes (obtained by the low-level features - mainly MFCC) series extracted from each stream sub-segment and the phonetic representation supplied by the lexical model

Speech recognition can be carried out with different word error rates depending on the use of adaptation process for the acoustical and speaker models. This phase is commonly called adaptive (or “second pass”) speech recognition. The results of the non-adaptive phase (first pass) are used for training adaptive models.

Important factors that can influence the effectiveness of the adaptive phase can be identified in acoustic environment speech conditions (e.g. noisy speech), speaker overlapping and speech quality (e.g. planned vs. spontaneous speech)

Overall computation costs: very high – it requires a lot of memory for the models (1,5-2 Gbyte of RAM are required) and a lot CPU time (0,3x slower than real-time) in order to perform a single-pass speech to text process. Adaptive phase doubles the time requirement (thus adaptive pass can be executed only after the non-adaptive pass)

Any indication about computational costs refers to an actual off-the-shelf system (dual Pentium@2,8GHz).

Part C: Joint Audiovisual Content Analysis and Structuring Tools

Techniques for structuring, classifying and summarising audiovisual content are useful for facilitating documentation and presenting retrieval results. Such techniques include for example shot type classification, extraction of representative items (e.g. key frames), scene classification or summary generation. Tools based on different audiovisual features as well as on fusion of visual, audio and speech analysis results will be surveyed.

14 Scene/Story Segmentation

Like shot boundary detection, scene or story segmentation aim at temporally decomposing the video into coherent units. There are however two fundamental differences:

- A shot is a well defined concept, thus the task of detecting shot boundaries is well defined. In contrast, the definition of a story or a scene is not unambiguous and may be genre dependent. For example, it is easier to define a story in a news broadcast than a scene in a feature film.
- Shots directly depend on low-level features. They can thus be derived directly from the visual signal. Stories and scenes are high-level concepts, defined by their semantics. Detecting scenes and stories thus requires bridging the so called semantic gap between low-level features and semantic concepts.
- Shots are a purely visual property. Audio features may continue unchanged at a shot boundary. As scenes and stories are high-level concepts, they depend on all modalities.

The assumption, that scene or story boundaries coincide with shot boundaries (i.e. that scenes and stories are groups of shots) is commonly used, as it simplifies the problem. However, this assumption does not always hold, as scene or story boundaries also appear within a shot (e.g. a shot in which a new programme anchorperson who finished one topic and starts talking about another).

14.1 Scene/Story Definition

The problem defined as scene/story segmentation can be regarded as the process of identifying the parts, or sub-units, starting from the whole represented by an audiovisual content.

The very first attempt at characterising what is a unity in the context of art works can be traced back to the formulation of the so called Aristotelian units. In general, that characterisation dealt with the definition of a set of best practices for the production of theatre operas, and it has been developed and improved by Renaissance scholars. These principles were the unity of action, of place and of time.

Having collected in the past centuries both a vast success, as well as many points of disagreement, from the point of view of modernity, these attempts reveal insufficient in accounting for the complexity of the language and expressions that nowadays permeate artistic production in general and audiovisual content production in particular.

Besides, if dealing with a posteriori analysis of merely audiovisual content, it is hard, if even logically impossible, in general, to reconstruct the original intentions of the authors of a work unless some form of a priori knowledge can be used.

Despite these aspects, there have been various attempts to achieve precise definitions. The term logical story unit (LSU) has been proposed [Hanj99b][Hanj00] as an approximation of a movie episode. The authors describe a LSU as a sequence of temporally contiguous shots, that contains groups of shots which are linked by visually similar content elements. [Vend02] refine the definition by defining the LSU based on its boundaries, which correspond to perceived boundaries in place and/or time. Shots within the same LSU are visually similar, those of different LSUs are visually dissimilar.

In [Sund02] the authors introduce the concept of computable scenes, which is a scene definition based on low-level visual features and thus can be unambiguously derived from the video data. The main property of a computable scene is long term consistency in both visual and audio signal.

Hanjalic [Hanj04] extends this, by defining two conditions for automatic scene/story segmentation, i.e. for finding changes in the coherence of content:

- Computability: The existence of a feature set that is capable of revealing changes in the content coherence.
- Parsability: If the content has been generated as a concatenation of semantic segments, and if the content coherence is computable, the content is parsable.

This work also contains an overview of a number of approaches to content segmentation.

14.2 Approaches

The approaches to scene and story segmentation can be classified into approaches, which are based on prior knowledge about the structure of a programme in a certain domain (such as the appearance of an anchorperson), and into those using a statistical framework to classify that is being trained on sample data. In [Hsu04] the approaches are called heuristic rule based and statistical approaches, in [Li03] the authors make a similar distinction into model-based and non model-based approaches.

However, when looking at the algorithms proposed in literature, it seems that the most distinguishing feature is the application area. Most of the work is about news story segmentation, only a part deals with sports programmes (interestingly most of the work about sports programmes is model-based) and feature films. We thus use the target application as structuring criteria for this section.

14.2.1 News Story Segmentation

News story segmentation is the most common application of scene/story boundary detection. Many of the approaches use a combination of visual, audio and speech features for the segmentation. The features used by different approaches are often similar.

The basic features employed in many of the approaches are visual similarity between shots within a time window and the temporal distance between shots, e.g. [OCon01][Haup98][Eich04][Hoas04]. Some approaches use additionally the similarity of faces appearing in the shots [Haup98].

Some methods rely on some prior knowledge about the structure of the news broadcast, such as the repeated appearance of anchor person shots. There are basically two approaches: classifying shots as anchor person shots because of certain properties such as faces [Hsu04][Pick03], or by using a key frame of an anchor person shot as query of a similarity search [Zhai04][Volk04].

The audio signal is used in many approaches to detect pauses, which may indicate a topic change [Hsu04][Haup98][Volk04][Que04][Hoas04][Eich04]. Other audio features which are used are speaker change [Hsu04][Chai02][Que04], change between music and speech [Que04][Chai02][Hoas04], detection of jingles [Que04] and changes in the acoustic environment, such as changes of the SNR (to discriminate studio from outside shots) [Haup98].

A number of approaches also use text from transcripts or automated speech recognition. Some use the text to find similar word appearances in different shots [Hsu04][Haup98] or watch for trigger phrases that indicate certain types of shots [Eich04]. There are also systems which rely heavily on text similarities between the shots [Pick03][Volk04].

Some of the reported approaches are based on supervised learning approaches. The work discussed in [Hsu04] and [Hoas04] is based on classifying boundary candidates into story boundaries and non story boundaries using the expectation-maximisation algorithm (EM) and support vector machines (SVMs) respectively. The approach in [Chai02] is based on shot classification using a hidden Markov model (HMM).

The approach reported in [Zhai04] uses a model called shot connectivity graph (SCG). Shots are classified and each node in the graph represents a shot, the edges are transitions from one shot to another. As it is expected that anchor person shots reappear, the task is to search for cycles in the graph. Special types of shots are detected using other features, such as word spotting for detecting sports shots and greenish/bluish colour impression to detect weather shots.

14.2.2 Segmentation of Feature Films

Most of the early general approaches on scene segmentation are based on visual features. The basic assumption is that there a groups of shots in a scene which are more similar in terms of colour properties than shots belonging to different scenes (cf. [Rui99], [Hanj99b]). To reduce the amount of images to be processed, key frames are used to represent the colour properties of a shot.

Recently these approaches have been extended by further features. Based on the approach by [Hanj99b], a scene definition based on three different types of events has been proposed in [Tava04]. The type of event is determined from visual properties of the key frames, from which only defined regions are used in order to be robust against motion in the centre of attention.

In both [Rash03] and [Truo02] use of scene dynamics (film tempo), i.e. the length of the shots and the amount of motion is proposed. To be robust against camera motion, the amount of local motion (i.e. object motion relative to the camera motion) can be used as feature. The work reported in [Truo02] also uses fades and dissolves as scene boundary candidates and refines the common colour similarity measure by using the difference by the most highly saturated colours of the shots.

Together with the proposal of the concept of computable scenes (cf. 14.1), an approach for the detection of the boundaries of computable scenes is proposed in [Sund02]. In this work scene boundary candidates are extracted from both the visual and audio signal, and the candidate sets are then fused. The main features used are chromaticity and lighting similarity between the key frames, correlation in the audio signal and silence. The boundaries are determined as minima of a coherence function between the shots. A topology graph is built to detect patterns of shot sequences (such as dialog scenes).

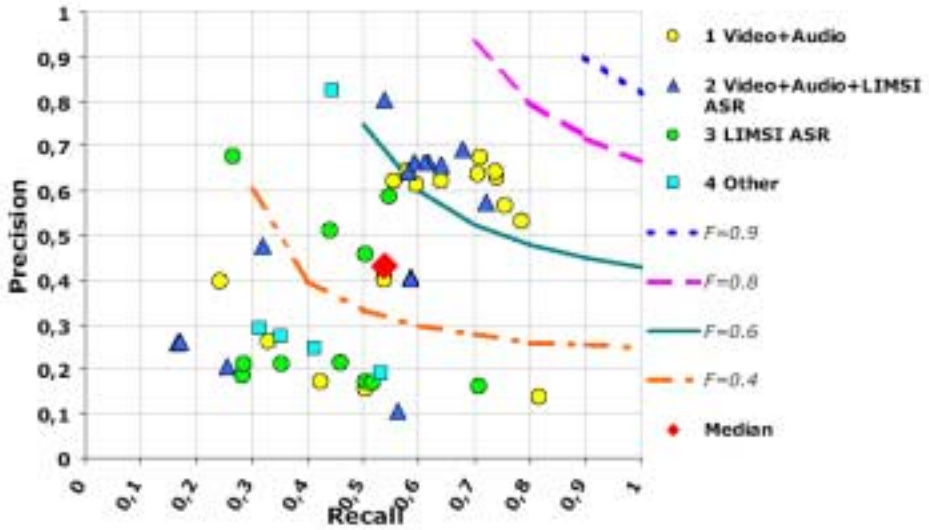


Figure 4: Precision and recall measures of news story boundary segmentation. The results have been generated by using different modalities as input (1) video and audio signal, (2) video and audio signal and speech to text, (3) only speech to text (4) other combinations [Kraa04].

14.2.3 Segmentation of Sports Programmes

In [Nitt02] a model-based scene segmentation approach for sports programmes is proposed. The segmentation is mainly based on speaker changes in the speech signal. The closed caption text is used to classify the segments to one of a set of types using a Bayesian network. The model, that consists the possible sequences of these segments is then used to infer the higher level structure.

The approach reported in [Zhong01] is also model-based, but uses mainly visual features to classify shots to certain event types. The features used are colour, the object layout (i.e. the position of moving objects), and the edge structure (lines of the field).

14.3 Performance of News Story Segmentation

Although evaluation methods for story segmentation have been proposed (cf. [Vend02]), the performance of the different approaches is difficult to compare and evaluate. Since 2003 a news story segmentation task is part of the TREC Video Retrieval Evaluation [Trecvid]. The advantage of news broadcasts is that they have a clear story structure.

Figure 4 shows the results that can be achieved with state of the art news story segmentation algorithms. It also illustrates the dependency of the result on the modalities used as input. The joint use of video and audio features provides generally better results, while the additional use of speech recognition seems to slightly improve precision, but not recall.

Scenes which have been mis-classified by all or many of the systems and their properties are discussed in [Kraa04].

15 Shot and Scene Classification

Classification is done on different levels of granularity, from single shots to whole programmes. A common application is genre classification of scenes or programmes, for example to detect commercials or discriminate between news and feature films.

Another group of approaches deals with classifying shots or scenes by sites or objects, e.g. indoor/outdoor, natural/man-made environment or the presence of certain objects. These approaches are also referred to as concept detection. The term concept often includes not only sites and objects, but also events, for which a more detailed description can be found in Section 16.

Other algorithms aim at labelling or categorising shots and scenes, for example, by the type of view in a sports broadcast, or by matching scenes of a movie shot at the same location or setting.

15.1 Genre Classification

The genre classification methods reported in literature work on different levels of granularity, from single shots to whole programmes. A summary of the properties of common types of genres can be found on [Snoek05].

One of the first genre classification algorithms is the work proposed in [Fis95]. It works in three steps, starting with a syntactic analysis of the video, then extracts style attributes and finally maps the style attributes to genres. It has been used for classification into the genres, news, car race, tennis, commercials and cartoons.

In [Truo00] videos are classified into sports, news, commercials, cartoons and music using a decision tree approach. The classification is based on the features clip length, percentage of fades, camera motion activity, changes of luminance, length of motion runs, colour coherence and percentage of high-brightness pixels.

The work reported in [Iane03] classifies videos into cartoons and non-cartoons.

The approach reported in [Goh04] is a commercial detection based on audio classification and motion activity. The shots are clustered based on their visual and audio features. The authors report that they have also successfully applied the algorithm to highlight detection in soccer video.

The work presented in [Sug03] is based on the idea that there are three basic scene types in a movie: dialogues with action, dialogs without action and action without dialogue. The algorithm thus classifies shots into action, dramatic, conversation and generic (everything else). The features used for classification are motion intensity, shot length, audio power and audio classification.

A video classification system based on the detection on overlaid text and faces is presented in [Dim00]. Based on the number of lines of text, and the number, size and position of faces, a shot is assigned on of 15 labels. From these labels, the genres news, commercial, sitcom and soap are inferred using Hidden Markov Models.

In [Rash02] an approach for classifying movie previews by their genre is presented. The motion activity and the shot length are used to discriminate into action and non-action movies. Action movies are

further analysed by their audio feature and histogram to detect fire and explosion scenes. Non-action movies are classified based on the lighting properties in comedy, horror and drama.

The approach in [Jin04] uses a number of audio features (those defined by the MPEG-7 audio descriptors), motion activity, camera motion and HSV histograms for genre classification. Using a decision tree approach, the input videos are classified into the genres cartoon, drama, music video, news and sports.

In [Gil04] an approach for classification of shots into the four genres sports, drama, scenery and news is presented. It is based on an 8 dimensional feature vector representing the temporal variations in the shot, called “activity power flow” and the motion intensity. The activity power flow is derived from in the MPEG compressed domain from the number of blocks with reliable motion vector, unreliable motion vectors or which are intra-frame coded. A radial base function network (RBF) is used for classification.

There are a number of approaches that only use audio features to classify audiovisual material, which are not considered here.

15.2 Concept Detection

In [Naph03] the concept of a probabilistic multimedia object, called multiject, is introduced. A multiject is a term denoting an object, a site or an event. The approach is based on colour, texture and shape features, from which multijects are inferred using a support vector machine (SVM, see [CriST00] for an introduction)). A network is defined that relates the multiject among each other, defining positive and negative relations (i.e. positive relation between “road” and “outdoors”, but negative between “road” and “indoors”).

In [Rau03] the use of self-organising maps (SOMs [Koho97]) has been proposed for semantic concept detection.

The approach proposed in [Souv04] is based on visual features extracted from keyframes, the motion activity of a shot and a text transcript. The shot classifications are assigned form a training sample using a k-NN classifier, a support vector machine and a specialised classifier for keywords, using a fusion step to combine the results of the three approaches.

The work reported in [Chen04] uses a learning Bayesian network for classification and event detection. The low-level features used are visual colour and texture features, audio FFT and MFCC (Mel Frequency Cepstrum Coefficients), motion energy and face detection and video OCR results. The low-level features are mapped to a set of 168 concepts using a support vector machine (SVM). Combinations of these concepts are mapped to the classes to be detected.

A similar approach is used in [Amir04], where also a SVM is used to map a set of low level features to a set of concepts. In a second step the classes to be detected are inferred from combinations of the concepts.

The work described in [Snoek04] introduces a method called Semantic Value Chain for concept detection. The three links in the chain are the Content Link, using the low-level visual, audio and text features, the Style Link and the Semantic Context Link. The Style Link uses four style detectors: layout (based on shot length, text overlays, silence and voice overlay), content (faces and their location, cars, object motion, named entities in overlaid text or speech), capture (camera distance and motion) and concept (reporter names, content link). The first two links rely on support vector machines (SVM), for the Semantic Context Link, context vectors and ontologies have been used. The approach uses a lexicon of 32 pre-defined concepts.

15.3 Labelling and Categorization

In [Wang04] a scene classification approach for soccer video is presented. Keywords are assigned based on visual and audio features are separately. The visual features are motion activity and colour similarity within segments having similar motion, the audio segments are classified. The keywords stemming from visual and audio analysis are fused using a support vector machine (SVM). A similar approach, that performs segmentation and classification of scenes in soccer video is described in [Sun03].

An approach for classifying the scenes in a soccer video is reported in [Pap04]. The approach uses an alphabet of scenes, which are described by the pitch content in the frame. A hidden Markov model is used for classification.

In [Dor04] a framework for scene classification is presented. It uses a support vector machine which is trained by relevance feedback from the user. The framework has been applied to the classification of still images into the categories animal, city view, landscape and vegetation. As features the MPEG-7 descriptors colour structure, edge histogram and homogenous texture have been used. The performance on a set of 1200 images, of which 40% have been used for training, is between 74 and 87%.

An approach for classifying still images into indoor and outdoor scenes is presented in [Ser02]. As features a colour histogram computed in the LST colour space and texture features calculated from Daubechies wavelets are used. The image is divided into 4x4 sub-blocks and the features are extracted for each sub-block. The classification is done in two stages, both are using support vector machines (SVMs). In the first stage colour and texture features are treated separately, while in the second step the decision is done based on the indoor/outdoor likelihoods resulting from colour- and texture-only classification.

In [Szum98], an approach to indoor/outdoor classification of images is presented. The features used are a colour histogram in the Ohta colour space (the colour axes in this space are found as the principal components in the RBG space of a large set of natural images), MSAR texture features and a frequency feature calculated from a DFT with subsequent DCT. The authors use a k-nearest neighbour classifier and report a performance of 90% on the test set.

The work in [Vail98] is an approach to binary image classification problems. The authors propose the approach for classification into city and landscape images, and the extend the approach to classification into sunset/sunrise, forest and mountain images by first classifying into sunset/sunrise and forest/mountain, and then split the second category in a subsequent classification step. The features used are colour histogram, colour coherence vector, DCT coefficients, edge direction histogram and edge direction coherence vector. The authors report a performance of over 90% for both classification tasks. A similar approach is used in [Vail00] to detect sky and vegetation in images by classifying blocks of the image.

The work described in [Scha02] aims at matching scenes of a video that have been shot at the same scene or show the same rigid main object. The approach uses the key frames of the video, which are described by invariant descriptors. The descriptors of different key frames are then matched using various steps of checks ranging from local to global.

The work on image classification reported in [Guy02] is based on human classification. In an experiment, users selected the most similar out of 8 images w.r.t. a sample image and assigned a degree of similarity between the images. From the collected data, a global distance matrix of the total set of 105 images and projected to a 2D space using Curvilinear Component Analysis (CCA). This data is used to build a model of human classification, which is related to image low-level features, which are extracted using a bank of Gabor filters.

15.4 Affective Content Analysis

In [Hanj04] the distinction between content analysis on a cognitive and on an affective level is made. All the classification criteria discussed above, such as genre classification or detection of settings or objects work on a cognitive level.

Apart from that, the content has an affective level. Affective content analysis could for example aim at finding the most funny scenes of a comedy or the most thrilling scenes of a horror movie. Of course the affective reaction is subjective, but Hanjalic argues that there is an affective reaction that is expected by the creator of the content, which is based on general experience and typical reactions of the audience. Thus, content analysis can aim to extract the expected affective response.

Hanjalic selects three dimensions of the affective response: valence (pleasant/unpleasant), arousal (calm/excited) and control or dominance (controlled/uncontrolled). He proposes algorithms to extract the values in the three dimensions from a set of low-level visual and audio features.

15.5 Performance of Concept Detection Algorithms

Many of the features of the TREC Video Retrieval Evaluation [Trecvid] feature extraction task are related concept detection or classification of video shots or segments. In TRECVID 2004 these were the following features [TVKA04]:

- Boats/ships
- Trains
- Beach
- Road

Additionally, the following features from TRECVID 2003 are of interest in this context [TVKA04]:

- Indoor
- People
- Building
- Vegetation
- Non-studio setting
- Sporting event
- Weather news

From the result in Figure 5 and Figure 6 can be seen, that the results are very different for the various features. The precision for some is good enough for practical use (e.g. weather news, sporting event). The reason is that are linked more clearly to certain low-level features, e.g. weather news has characteristic colours, sporting events have high motion energy and a high background noise level. For concepts where this relation is not so well defined, the precision is very poor (e.g. beach).

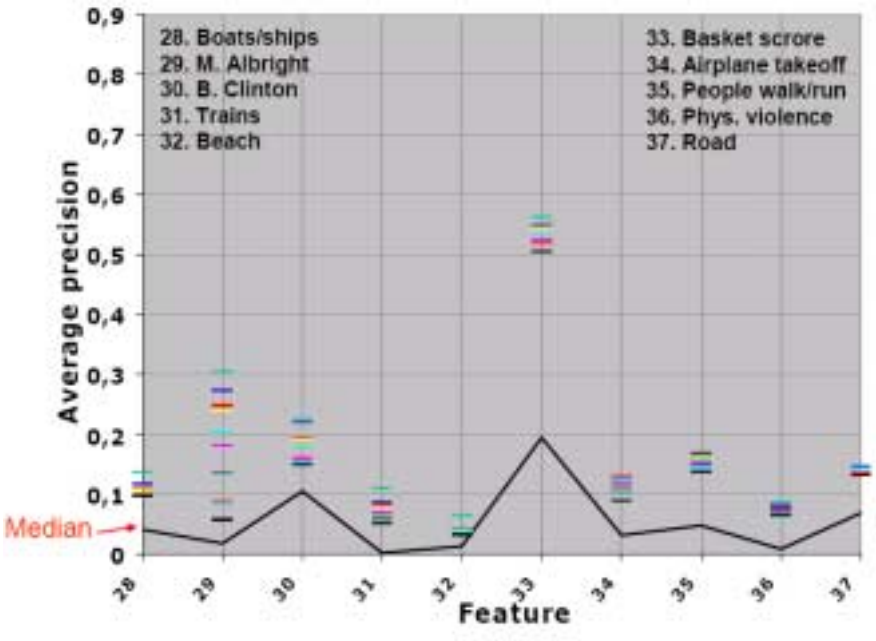


Figure 5: Average precision of feature extraction in TRECVID 2004. The stacks of short lines represent the top 10 runs, the thick continuous line is the median [TVKA04].

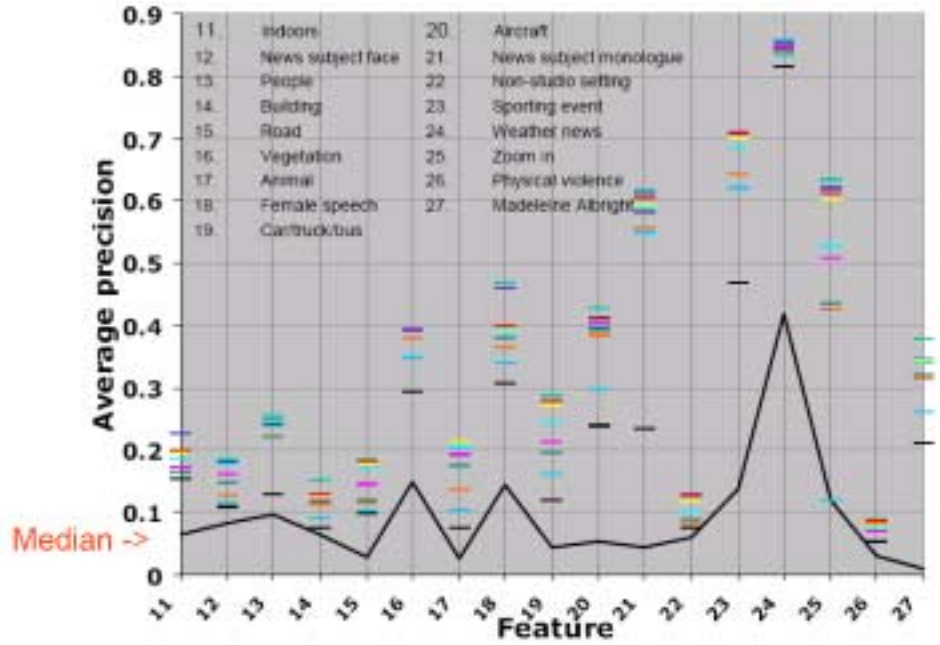


Figure 6: Average precision of feature extraction in TRECVID 2003. The stacks of short lines represent the top 10 runs, the thick continuous line is the median [TVKA04].

16 Event Detection

Event detection deals with locating segments of audiovisual content, that contain a relevant action, such as a dialog, a goal in soccer game etc. A fundamental problem is that there is no generally applicable definition of an event – it is very content specific what constitutes an “interesting” event in audiovisual content. Event detection is related to classification; some event detection approaches rely on classification of shots before event detection is done based only on the classification results.

A big part of the literature deals with sports video, because, as pointed out in [Ada03], they have a well defined structure with a limited set of events, and it is a priori clear which events are relevant.

Another common type of events are dialogs, as they are important features in both news and feature films. There is a class of algorithms that only deals with the detection of dialog scenes.

Some researchers have tried to find generic approaches to the problem, which are not specific to the type of events or dependent on prior knowledge about the domain.

There is a considerable amount of work on event detection for surveillance application. However, as many of the reported approaches are rely on conditions, which do not hold for content analysis of audiovisual broadcast and film material, such as the assumption of a static camera, these approaches have not been considered here.

16.1 Event Detection for Sports Video

The majority of the event detection literature deals with events in sports video. A good overview on the approaches and the different sports covered can be found in [Ada03]. Many of the methods rely on domain knowledge, i.e. on prior knowledge about the audiovisual features related to the events (e.g. the crowd will cheer after a goal).

In [Leo04] a framework for the detection of goal shot sequences in soccer games is presented. The features used are the motion properties, the fact that goal shots are usually shown in at least two different shots and the loudness of the audio signal. The results show that using only the visual features is nearly as good as using both audio and video, while the audio features alone give weak results.

The work presented in [Ber04] also deals with soccer and is based on the observation that the camera motion almost always corresponds to the motion of the ball. Using camera motion information and the lines that are visible in the playfield, events taking place in the penalty areas can be detected.

The approach in [Xio03] is only audio based and can be used for different sports. The prior assumption about the event is that there is applause and cheering after the event. The features used are the background noise level and the classification of segments of the audio signal into applause, speech and music (using a Hidden Markov model).

Like the previous approach, also the work reported by [Sad04] is applicable to a larger number of sports, namely all field sports. Features used for detection are the amount of grass hue, the angles of the lines of the playfield, close ups, motion activity, detection of images showing crowds (high frequency content over large parts of the image) and the activity in the speech signal. The event detection algorithm is based on a support vector machine.

The approach proposed for event detection in soccer video in [Kang03] is based on classification of segments. Based on the classification, visual keywords are assigned. The event detection is done by inferring events from the assigned keywords. An improvement of this approach, with separate classification of visual and audio data and an additional fusion step has been proposed in [Wang04].

In contrast to other approaches, the work in [Pan01] is based on common editing rules. The assumption is that there is a slow motion replay segment after an important event. A Hidden Markov model is used to detect candidates of replay segments. To discriminate them from commercials, further conditions, such as colour similarity to other parts of the sports video, must be fulfilled.

In [Tov01] a framework based on the use of a semantic model of soccer events and rule based event detection is introduced. Players and ball are tracked and events are modelled as sequences of interactions among players or between players and the ball. The events are described using a hierarchical entity-relationship model.

16.2 Dialog Scene Detection

Dialogs are an important class of events in audiovisual material. A number of approaches thus deal with the detection of dialog scenes between two or multiple speakers. A more detailed version of the overview in this section can be found in [Hab04].

The existing papers and proposed systems are grouped into the following three basic architectures to provide a better overview:

- Scene Based Classification: First the video is segmented into scenes. Then the scenes are classified whether they contain dialogs or not.
- Direct Dialog Scene Detection: In a first step features for the shots are calculated. Then the dialog scenes are detected directly, e.g. using a HMM (Hidden Markov Model).
- Shot Based Classification: First the shots are classified whether it is probable that they are part of a dialog scene or not and then shots are grouped into dialog scenes.

16.2.1 Scene Based Classification

In [Li03] a system for segmenting videos into scenes and classifying these scenes into one of three classes (two-speaker dialogs, multiple-speaker dialogs and hybrid events) based on audiovisual information is proposed. After a shot detection step, the shots are grouped into scenes based on visual and acoustic similarity using Sink-based Scene Construction implemented as a Window-based Sweep Algorithm. The proposed window-based sweep algorithm in a first step determines the shot sinks by comparing the key frames of the shots within a window using a threshold for the visual similarity. In the next step, temporally overlapping sinks are grouped into scenes (referred to as events in the book). Subsequently, the scenes are classified by a set of heuristically derived rules.

16.2.2 Direct Dialog Scene Detection

In [AAW01] a whole system for audiovisual dialog scene detection is proposed based on acoustic information (classification of speech, silence and music) and visual information (face and location change information). The first step is shot boundary detection, and for each shot audio and visual

features are extracted. The resulting feature vector is classified by a classifier based on a Hidden Markov Model. In this paper only the classifier itself was implemented, all low level input data were annotated manually.

In [Ala02] a system for dialog scene detection similar to the one in [AAW01] is described where the low-level input is extracted automatically. Automatic feature extraction is usually less accurate than manually extraction—therefore the results are worse.

A System to detect dialog and action scenes that uses a top-down approach based on video editing rules is presented in [CRÖ03]. An audio classifier based on a Support Vector Machine is the input for a complex Finite State Machine which identifies dialog and action scenes. In the paper, the term action scene is used to address one-on-one fighting action scenes.

Saraceno and Leonardi [SL99] considered segmenting a video into the following basic scene types: dialogs, stories, actions, and generic. This is accomplished by first dividing a video into audio shots and visual shots independently and then grouping video shots so that audio and visual characteristics within each group follow predefined patterns. First, the audio signal is segmented into audio shots, and each audio shot is classified as silence, speech, music, or noise. Simultaneously, the video signal is divided into video shots based on colour information. For each detected video shot, typical block patterns in this shot are determined. Finally, the scene detector and characterisation unit groups a set of consecutive video shots into a scene if they match the visual and audio characteristics defined for a particular scene type. Then successive shots are compared and labelled sequentially: a shot that is close to a previous shot is given the same label as that shot; otherwise, it is given a new label.

Lienhart et al. [PLE99] proposed to use different criteria to segment a video: scenes with similar audio characteristics, scenes with similar settings, and dialogs. The scheme consists of four steps. First, video shot boundaries are detected. Then audio features, colour features, orientation features, and faces are extracted. Next, distances between every two video shots are calculated, with respect to each feature modality, to form a distance table. Finally, based on the calculated shot distance tables, video shots are merged. The authors also investigated how to merge the scene detection results obtained using different features. The authors argued that it is better to first perform scene segmentation/classification based on separate criteria and leave the merging task to a later stage that is application dependent.

To examine audio similarity, an audio feature vector is computed for each short audio clip, which includes the magnitude spectrum of the audio samples. A forecasting feature vector is also calculated at every instance using exponential smoothing of previous feature vectors. An audio shot boundary is detected by comparing the calculated feature vector with the forecasting vector. The prediction process is reset after a detected shot cut. All feature vectors of an audio shot describe the audio content of that shot. The distance between two shots is defined as the minimal distance between two audio feature vectors of the respective video shots. A scene in which audio characteristics are similar is noted as an audio sequence. It consists of all video shots so that every two shots are separated no more than a look ahead number (three) of shots and that the distance between these two shots is below a threshold. A dialog scene is detected by using face detection and face matching techniques. Faces are detected by modifying the algorithm developed by Rowley et al. [RBK98]. Similar faces are then grouped into face-based sets using the eigenface face recognition algorithm [TP91]. A consecutive set of video shots is identified as a dialog scene if alternating face-based sets occur in these shots. The above scene determination scheme has been applied to two full-length feature movies. On average, the accuracy of determining the dialog scenes is much higher than that for audio sequences and settings. This may be because the definition and extraction of dialog scenes conform more closely to the human perception of a dialog. The authors also attempted to combine the video shot clustering results obtained based on different criteria. The algorithm works by merging two clusters resulting from different criteria, whenever they overlap. This has yielded much better results than those obtained based on audio, colour, or orientation features separately.

16.2.3 Shot Based Classification

[SPSV01] proposes a Multi-Expert System based on the following ideas:

- A scene is a group of semantically correlated shots.
- Almost all shots belonging to a dialog scene can be characterised as dialog shots.
- Shots belonging to the same dialog scene are temporally adjacent.

Similarly to the systems proposed in [AAW01][Ala02] the first step in the detection process is shot boundary detection. In the next step, three experts (Face Detection Expert, Camera Motion Estimation Expert and Audio Classification Expert) classify the shots. Then the output of the experts is combined using a combining rule to classify each shot as (dialog / no dialog). Subsequently, the classified shots are grouped into dialog scenes using a Finite State Machine with three states (dialog scene, probable dialog scene and not dialog scene).

16.3 Generic Approaches

In [Zel01] an event is defined as a temporal object, which is characterised by features over multiple temporal scales. The authors thus propose the extraction of a “temporal texture” from a temporal pyramidal representation of the image sequence (i.e. sequences having different temporal resolution). The distance between segments is calculated as the normalised difference between the temporal texture features of the segments. Events are found as the results of clustering by temporal texture. The approach is not based on any prior knowledge.

The work in [Rui00] detects candidates of event boundaries by detecting discontinuities in the motion pattern of the sequence. A sequence of optical flow fields is estimated and the background motion is subtracted. The singular value decomposition (SVD) is used to determine the salient motion vectors in each of the fields and the trajectories of the SVD coefficients are analysed to detect discontinuities. The resulting set of discontinuities is a superset of the set of event boundaries, i.e. the algorithm results in temporal oversegmentation.

The algorithm proposed in [Hae00] and discussed in more detail in [Hae01] is generic in most of its steps, only two of the steps require domain knowledge. After global motion compensation, motion blobs, i.e. image regions that are not only subject to the global motion, are detected and their centres are estimated. Colour and texture descriptions of the blobs are extracted, and the spatial and spatio-temporal relations between the blobs are described. The blobs are then classified using a back-propagation neural network, the classes are of course domain specific. In the final step, events are inferred from the blobs, their features and relations using domain specific rules. The author report that the algorithm has been applied to detecting hunt sequences in wildlife video, as well as to the detection of plane landings and rocket launches.

The work reported in [Leh04] aims at detecting action sequences in motion pictures. The features used are the shot length, the motion activity in the shot and the activity of the camera motion. The detection of action sequences is based on editorial principles for action scenes, such as short shots or fast camera motion.

In [Nam98], an approach for detecting violent scenes in movies is presented. The features used are the spatio-temporal dynamic activity, the detection on flames and explosions based on colours, the change of the number of blood colour pixels, and the sound effects (shots, explosion sounds).

16.4 Performance of Event Detection Algorithms

Some of features of the TREC Video Retrieval Evaluation [Trecvid] feature extraction task related to event detection. In TRECVID 2004 these were the following features [TVKA04]:

- Basket score
- Airplane taking off
- People walking
- Physical violence between people

The results show that the algorithms used are getting less specialised. However, as can be seen from Figure 7, the performance is quite different for different types of features and generally the precision that can be achieved is limited.

The algorithms used in TRECVID 2004 have been discussed in the previous subsections or in Section 15, if the same algorithm has been used for both classification and event detection.

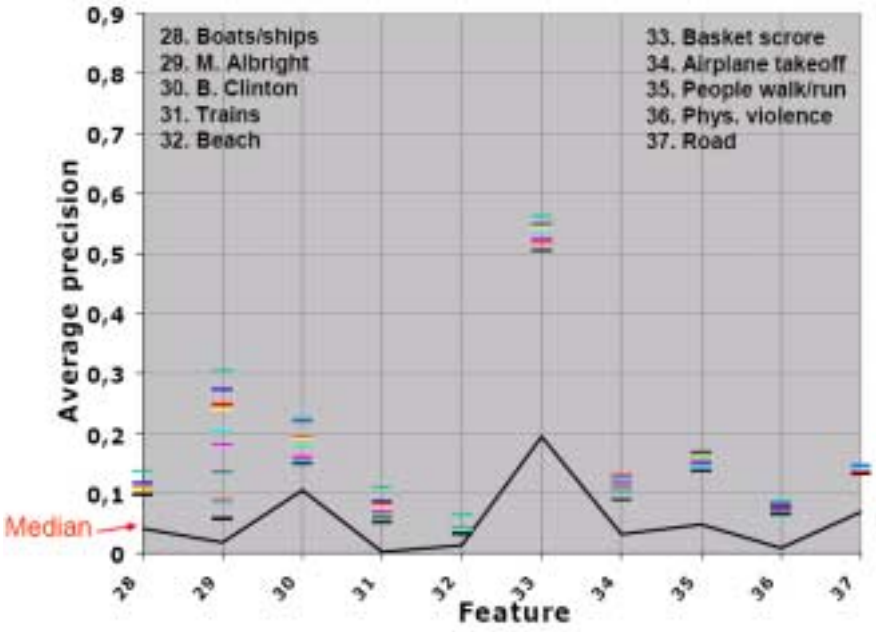


Figure 7: Average precision of feature extraction in TRECVID 2004. The stacks of short lines represent the top 10 runs, the thick continuous line is the median. Features 33-36 are relevant for event detection [TVKA04].

17 Video Content Abstraction

17.1 Motivation

Having a presumably long (e.g. 250min) video, one can not tell what it is about without watching the whole clip. If it is possible to abstract the main message of the movie - that is give an overview about events, actors, etc. - it is possible to get a glance of its content in a short period of time. This paper examines techniques to produce such overviews based on [Ying03] and [Ying01].

17.2 Scope

This report focuses on the state-of-the-art of video content abstraction, i.e. to give an overview of some audio-visual content. Basically one can think of two types of abstraction:

- textual abstraction, i.e. give a textual representation of a clip (in natural language)
- visual abstraction, i.e. give a visual representation of a clip

As it turns out, the first option is quite hard to realise, though for specific domains and applications some proposals do exist [Gerb02], [m4]. This paper therefore focuses on visual-based abstraction methods.

In the field of visual abstraction, there exist primary two different types of abstracts: still-image and moving-image abstracts. The still-image abstract is a collection of images extracted from the original footage, while the moving-image abstract is a video clip of considerably shorter length containing distinct and/or important events, actors and so on.

Research community has coined the expression video summary for still-image based abstracts and video skim for moving-image abstracts. We will stick with this kind of naming convention throughout this paper.

There are some significant differences between video summary and video skimming. A video summary can be built fast, since generally only visual information is utilised and no handling of audio and textual

information is needed. Therefore, once composed, it is displayed more easily since there are no synchronisation issues. On the other side video skimming offers more natural approach w.r.t. users, as they can watch a trailer compared to a slide show - finally in many cases motion also bears information.

In theory video abstracts can be produced both manually and automatically, but due to the sheer volume of content, fully automated video analysis tools play a more and more important role.

17.3 Video Summary

There is a lot of research work going on in the video summary area compared to video skimming.

Because the video summary is existentially a collection of still images that best represent the underlying content, the extraction or generation of these still-images (often called keyframes) becomes the main focus of all summary work.

17.3.1 Sampling-based Keyframe Extraction

Most of the earlier work in video summarisation chose to select key frames randomly or by uniformly sampling the video frames at certain time intervals, see e.g. the Video Magnifier [Mills92] system. The main drawback of this approach is that it may cause some short yet important segments to have no representative frames while longer segments might have multiple frames with similar content, thus failing to represent the actual video content.

17.3.2 Shot-based Keyframe Extraction

Another approach is to extract keyframes by adapting to the dynamic video content.

Since a shot is defined as a video segment within a continuous capture period, a straightforward way of keyframe extraction is to use the first frame of each shot as its keyframe, see e.g. [Ueda93], [Arm94] and [Eng96]. While being sufficient for stationary shots, a single keyframe per shot does not provide an acceptable representation of dynamic visual content. Hence multiple keyframes need to be extracted, though most of existing work chooses to interpret the content by employing some low-level visual features such as colour and motion, instead of performing an ample semantic analysis. The following shot-based approaches will be examined in the subsequent sections: colour-based approach, motion-based approach, mosaic-based approach and others.

17.3.2.1 Colour-based Approach

Zhang et al. reported in [Zhang97] a colour-based approach where keyframes are extracted in a sequential fashion for each shot. Particularly, the first frame within the shot is always chosen as the first keyframe, and then the colour-histogram difference between the subsequent frames and the latest keyframe is computed. Once the difference exceeds a certain threshold, a new keyframe will be declared. One possible problem with this method is that there exists a certain probability that the first frame is a part of transition effect at the shot boundary, therefore reducing its representative quality. Due to the invariance of colour histogram w.r.t image orientations and robust to background noises, colour-based keyframe extraction algorithms have been widely used. However, most of these works are heavily threshold-dependent, and cannot well capture the underlying dynamics when there is lots of camera or object motion.

17.3.2.2 Motion-based Approach

Motion-based approaches are relatively better suited for controlling the number of frames based on temporal dynamics in the scene. In general optical flow computation [Wolf96] or pixel-based image differences [Lag96] are commonly used in this approach. A domain specific keyframe selection method is proposed in [Ju98] where a summary is generated for video-taped presentations. Sophisticated global motion and gesture analysis algorithms are developed.

17.3.2.3 Mosaic-based Approach

A limitation of above approaches is that it is not always possible to select the keyframes that can represent the entire video content well. For example, given a camera panning/tilting sequence, even if multiple keyframes are selected, the underlying dynamics still couldn't be well captured. In this case, the mosaic-based approach can be employed to generate a synthesised panoramic image that can represent the entire content in an intuitive manner. Mosaics (also known as salient stills, video sprites or video layers), are – according to [Vasc98] - usually generated in the following two steps:

1. Fitting a global motion model to the motion between each pair of successive frames;
2. Compositing the images into a single panoramic image by warping the images with the estimated camera parameters.

The MPEG-7 MDS document [MPEG7-5] lists some commonly used motion models including translational model, rotation/scaling model, affine model, planar perspective model and quadratic model.

17.3.2.4 Other Approaches

Some other work integrates certain mathematical methodologies into the summarisation process based on low-level features. In the work reported by Doulamis et al. [Doul00], several descriptors are first extracted from each video frame by applying a segmentation algorithm to both colour and motion domains, which forms the feature vector. Then all frames' feature vectors in the shot are gathered to form a curve in a high-dimensional feature space. Finally, keyframes are extracted by estimating appropriate curve points that characterise the feature trajectory, where the curvature is measured based on the magnitude of the second derivative of the feature vectors with respect to time.

In [Stef00], Stefanidis et al. present an approach to summarise video datasets by analysing the trajectories of contained objects. Basically, critical points on the trajectory that best describe the object behaviour during that segment are identified and subsequently used to extract the keyframes. The Self-Organising Map (SOM) technique is used to identify the trajectory nodes.

Jung-Rim Kim et al propose in [Kim02] to use the notion of fidelity (as an attribute in MPEG-7 FDIS) that can be used for scalable hierarchical summarisation and search. The fidelity is the information on how well a parent key frame represents its child key frames in a tree-structured key frame hierarchy. In their paper, they demonstrate the use of fidelity for the summarisation of well-structured news by using temporal information as well as low-level features. A related approach was reported by Divakaran et al in [Div03].

17.3.3 Segment-based Keyframe Extraction

One major drawback of using one or more keyframes for each shot is that it does not scale up for long videos since scrolling through hundreds of images is still time-consuming, tedious and ineffective. Therefore, recently more and more people begin to work on higher-level video unit, which we call a video segment in this report. A video segment could be a scene, an event, or even the entire sequence. In this context, the segment-based keyframe set will surely become more concise than the shot-based keyframe set.

In [Uchi99], Uchihashi et al first cluster all video frames into a predefined number of clusters, and then the entire video is segmented by determining to which cluster the frames of a contiguous segment belong. Next an importance measure is computed for each segment based on its length and rarity, and all segments with their importance lower than a certain threshold will be discarded. The frame that is closest to the centre of each qualified segment is then extracted as the representative keyframe, with the image size proportional to its importance index. Finally, a frame-packing algorithm is proposed to efficiently pack the extracted frames into a pictorial summary.

Sun and Kankanhalli reported their work in [Sun00] where no shot detection is needed. On the contrary, the entire video sequence is first uniformly segmented into *L*-frame long units, and then a unit change value is computed for each unit, which equals to the distance between the first and last frame of the unit. Next, all the changes are sorted and classified into *2* clusters, the small-change cluster and the large change cluster, based on a predefined ratio *r*. Then for the units within the small-change cluster, the first and last frames are extracted as the R-frames, while for those in the large-change cluster, all frames are kept as the R-frames. Finally, if the desired number of keyframes has been obtained, the algorithm will stop, otherwise, the retained R-frames will be regrouped as a new video, and a new round of keyframe selection will be initialled. This work showed some interesting ideas, yet

the uniform segmentation and subsequent two-class clustering may be too coarse. A simple colour histogram-based distance between the first and last frame of a segment cannot truthfully reflect the variability of the underlying content, and if these two frames happen to have similar colour composition, even if this segment is quite complex, it will still be classified into the small-change cluster. Therefore, the final summarisation result may miss significant parts of the video information while at the same time retaining all the redundancies of other video parts.

Based on Lagendijk et al.'s work [Hanj97], Ratakonda et al. reported their work on generating a hierarchical video summarisation [Rata99] since a multilevel video summarisation will facilitate quick discovery of the video content and enable browsing interesting segments at various levels of details. Specifically, given a quota of total number of desired keyframes, each shot is first assigned a budget of allowable keyframes based on the total cumulative actions in that shot, which forms their finest-level summary. To achieve coarser-level summary, a pair-wise

K-means algorithm is applied to cluster temporally adjacent keyframes based on a predetermined compaction ratio *r*, where the number of iterations is controlled by certain stopping criterion, for instance, the amount of decrease in distortion, or a predetermined number of iteration steps. While this algorithm does produce a hierarchical summary, the temporal-constrained K-means clustering will not be able to merge two frames when they are visually similar but temporally apart. In some cases, a standard K-means will work better when preserving original temporal order is not required.

Compared with all above work, Dementhon et al. treat the video summarisation task in a more mathematical way where a video sequence is represented as a curve in a high-dimensional feature space [Dem98].

17.3.4 Other Keyframe Extraction Work

Other keyframe extraction work integrates some other technologies into their summarisation framework, such as wavelet transform, face detection, etc. In Dufaux's work [Duf00], he integrates the motion and spatial activity analysis with skin-colour and face detection technologies, so that the selected keyframe will have high likeliness of containing people or portraits, which certainly makes more sense than a landscape image. In Campisi et al.'s work [Camp99], a progressive multi-resolution keyframe extraction technique based on wavelet decomposition is proposed. One of the main advantages of this approach is the possibility of controlling the coarseness of the details' variations which are taken into account in the selection of a keyframe by properly choosing the particular sub-band to analyse and the level of the pyramid decomposition. However, the overall computation complexity is relatively too high.

Interesting results have been reported; however, this object-based summarisation scheme may only work well for videos which have relatively simple content and contain a small number of video objects, such as the surveillance videos. More complex video sources will reduce its effectiveness. A quite unconventional work using design patterns can be found in [Russ03].

17.4 Video Skimming

There are basically two types of video skimming: *summary sequence* and *highlight* [Hanj99a]. A summary sequence is used to provide users an impression about the entire video content, while a highlight only contains the most interesting parts of the original video, like a movie trailer that only shows some of the most attractive scenes without revealing the story's end.

Defining which video segments are the highlights is actually a very subjective process, and it is a very hard project to map human cognition into the automated abstraction process, thus most of existing video-skimming work focuses on the generation of a summary sequence. One of the most straightforward approaches in this case would be to compress the original video by speeding up the playback. As studied by Omoigui, et al. [Omo99], the entire video could be watched in a shorter amount of time by fast playback with almost no pitch distortion using the time compression technology. These techniques, however, only allow a maximum time compression of 1.5-2.5 depending on the speech speed [Hei86], beyond which the speech becomes incomprehensible.

The Informedia Project [Smi97] aims to create a very short synopsis of the original video by extracting the significant audio and video information. Particularly, text keywords are first extracted from manual transcript and closed captioning by using the well-known Term-Frequency- Inverse Document Frequency technique, then the audio skimming is created by extracting the audio segments corresponding to the selected keywords as well as including some of their neighbouring segments for

better comprehension. Next, the image skimming is created by selecting the video frames which are: a) frames with faces or texts; b) static frames following camera motion; c) frames with camera motion and human faces or text, and d) frames at the beginning of a video scene, with a descending priority. As a result, a set of video frames, which may not align with the audio in time, but may be more appropriate for image skimming in visual aspect are extracted. Finally the video skimming is generated by analysing the word relevance and the structure of the prioritised audio and image skimming. Experiments of this skimming approach have shown impressive results on limited types of documentary video that have very explicit speech or text contents.

In [Tok00], Toklu and Liou reported their work on video skimming where multiple cues are employed including visual, audio and text information. Specifically, they first group detected shots into story units based on detected “change of speaker” and “change of subject” markers that are sometimes available from the closed captioning. Then audio segments corresponding to all generated story units are extracted and aligned with the summarised closed-caption texts. Representative images are also extracted for each story unit from a set of keyframes consisting of all first frames of the underlying shots. Their final video skimming includes the audio and text information, but somehow the keyframe information is excluded from it. Finally, users are allowed to return their feedback to the system through an interactive interface so as to adjust the generated video skimming to their satisfaction.

Some other work in this area tries to find solutions for domain-specific videos where special features can be employed. The VidSum project uses a presentation structure, which is particularly designed for their regular weekly forum, to assist in mapping low-level signal events onto semantically meaningful events that can be used in the assembly of the summary [Russ00].

Another work reported by Lienhart mainly focuses on the summarisation of home videos [Lien00] where it is more usage model-based than content-based. First, the time and date information of the recordings are obtained by either extracting them from the S-VHS using text segmentation and recognition algorithms, or by directly accessing them from the digital video sequence. Then, all shots are clustered into five different levels based on the date and time they are taken, which include: the individual shots; a sequence of contiguous actions where the temporal distance between shots are within five minutes; a sequence of contiguous activities where temporal distance between shots are within one hour; individual days and individual multi-day events. In the next step, a shot shortening process is performed where longer shots are uniformly segmented into 2-minute-long clips. To choose the desired clips, the sound pressure level of the audio signal is calculated and employed in the selection process based on the observation that during important events, the sound is usually more clearly audible over a long period of time than is the case with less important content. Finally, all selected clips are assembled to form the final abstract using pre-designed video transition effects.

More recent approaches reported by Shi Lu et al using techniques based on graph optimisation can be found in [Lu04a] and [Lu04b].

Part D: Conclusion

18 Feasibility of Content Analysis Tools

This section summarises the content-analysis tools which have been discussed in detail in the previous chapters and draws conclusions concerning the practical usability of these tools inside the PrestoSpace metadata access and delivery factory.

18.1 Introduction

As can be concluded from this state of the art report, the scientific and technical literature in the field of audiovisual content analysis is quite extensive and vividly growing during these years. The problem that is now appearing is why to use these tools, which of them to adopt and where in a metadata factory system such as the PrestoSpace Metadata Access and Delivery factory.

The criteria with which to achieve this kind of assessments are not yet clear; however, what can be pointed out in form of very general indications is that all the algorithms and practices that have been studied and optimised, share some common aspects that can be summarised as follows:

- They are information extraction and recovering tools, i.e. their goal is to produce or reconstruct relevant pieces of information from the analysis of raw audio and video digital signals. The extracted/recovered information may be structural, i.e. regarding the mereological aspect of content, and/or semantic, i.e. regarding the concepts and situations expressed by means of images and sounds.
- Some of them are application- and context- dependant. To carry out their work, analysis processes need to have some amount of contextual information "wired inside" in some explicit or implicit way. The extracted information is always functional to the solution of determined problems in wider contexts.
- They are based substantially on some statistical multilevel analysis of a selected set of low-level features extracted directly from the audiovisual content. The set of selected features depends on the particular task to be solved and typically relies on the determination of their deemed relevance to human observers in the particular problem that is being tackled.

Quite obviously, in an ideal world where infinite manpower was available, any of the jobs carried out by whichever of the described tools could be performed by (yet good wish-endowed) humans with probably peerless quality, precision and efficiency.

Therefore, the principal rationale for the employment of these inventions is to be found in the dramatic lowering of needed resources that they allow yet obtaining sufficiently acceptable results for the purposes of a given application. This lowering should be ideally measured relatively to the resources employed by the overall value-adding process inside the metadata factory.

The above considerations enlighten how the available audiovisual content analysis tools should be preponderantly regarded as aiding supports for the documentation and characterisation of content finalised to the classification and subsequent automatic retrieval of audiovisual content.

As a conclusive concept, a correct path towards a sensible selection of which tools could be fruitfully adopted in the context of the present project should start from clear statements about what kind of content characterisation is thought useful to be done *automatically*. These guiding considerations should carefully take into account the levels of accuracy and efficiency that these mechanisms can afford in solving their tasks.

Accuracy is very well estimated by the levels of precision and recall showed by the tools. As for efficiency, useful parameters can be the process delay, the ratio between the media length and the processing time, the algorithms' class of complexity with respect to media duration, etc.

18.2 Influences to/from other areas

In the particular environment of PrestoSpace MAD factory, some interesting synergies could come up from the interactions with the other areas. Peculiarly this is evident with regards to the legacy metadata and to the semantic analysis area.

Legacy metadata could offer highly valuable inputs to the contextualisation of content analysis tools, for instance in terms of programme-level indication of occurring places and people, genre and content pre-classification, language information.

Semantic analysis could, on the other side, improve the efficacy and precision of content analysis tools by providing time information of occurring events and concepts extracted from the semantic analysis of pure text.

In the other way around, also results from the content analysis area can give support to other areas as, for example, that of semantic analysis.

18.3 Visual Content Analysis Tools

18.3.1 Low-level Visual Features

A large number of descriptors for low-lever features has been proposed throughout the last decade, a lot of this work comes from the context of content-based image/video retrieval (CBIR/CBVR). Many of these descriptors are well investigated. The visual part of MPEG-7 defined standardised descriptors for low-level features.

The low-level visual features colour, texture, shape and motion are important pre-requisites for all further mid- and high-level content analysis steps. Furthermore, they can be used for similarity search (query by example) within a video or over a collection of images and videos.

18.3.2 Spatial/Spatiotemporal Segmentation

Spatial and spatio-temporal segmentation are pre-requisites to extract regions and objects from image sequences. While spatial segmentation, e.g. colour and texture segmentation is a well investigated problem and reliable algorithms exist, spatio-temporal segmentation, using motion or a combination of motion with other features (e.g. colour) remains a hard problem. This is due to the difficulty of reliably reconstructing object motion from an image sequence.

Furthermore, segmentation of more complex constructs than just uniformly coloured or textured regions is only possible if other features, such as motion, can be used. This means that regions can be clustered because of their motion to segment a moving object, while segmenting a non-moving object would require general object recognition, which is still an unsolved problem.

18.3.3 Shot Boundary Detection

As far as hard cuts are concerned, shot boundary detection is a well-investigated problem and a number of reliable approaches exist. Detection of gradual transitions, such as fades dissolves and wipes, is still not completely solved, although a lot of progress has been made recently, as the results of the TRECVID 2004 shot boundary task show.

18.3.4 Video OCR

Text localisation and text segmentation in complex images and video have reached a high level of maturity. The same is true for OCR tools, although the conditions are harder on video data than in document OCR due to lower resolution and difficult background structures.

Currently, video OCR focuses on overlay text. Future research will also take scene text into account.

18.3.5 Face Detection and Recognition

Face detection and recognition approaches work satisfactorily in controlled environments, e.g. access control applications. Arbitrary backgrounds, settings an environmental conditions are severe problems for the algorithms. Because of the temporal redundancy and the availability of motion information, face detection and recognition work better on video than on still images. The results can be significantly improved, if reliable moving object segmentation is applied as a pre-processing step.

A lot of progress has been made in face recognition in recent years. However, in applications where the database of faces is large, the face features and descriptions turn out to be not sufficiently discriminative, which causes false matches. Recent approaches try to solve this problem by using 3D face models.

For face detection and recognition a considerable number of tools and software components are available, both commercial and non-commercial ones.

18.3.6 Defect and Quality Analysis

Defect and quality analysis and description of audiovisual archive content will enable quality based archive search functionality and will be a pre-requisite for efficient restoration and automatic restoration effort calculation. There are two major research areas in the field of quality and defect analysis corresponding to the measurement paradigm, the reference based and the non-reference based algorithms.

Reference based quality and defect measurement is widely used for determination of transmission channel quality, e.g. digital video broadcast or video conferencing. First normative (ANSI, ITU) solutions, e.g. for blocking, blur, edge business, noise are available. All these measures are based on the availability of undistorted reference essence and distorted encoded/transmitted/decoded essence. Although the availability of both is usually not the case for archived essence, algorithms and measures developed by these standards are an interesting base for research and development of reference free quality/defect measures.

Non-reference based analysis tries to find certain audiovisual quality/defect measures without the usage of any ground truth data. Because of this independence on ground truth data this approach has a wider application area than the reference based approach, e.g. essence available in an audiovisual archive very often requires reference free analysis. Because of the unavailability of ground truth data, calculation of quality/defect measures is more difficult than in the reference based approach. First non-reference based analysis approaches are available for the blocking, blurring, noise. Film/video medium originated defects analysis has been focussed until now on spatiotemporally very detailed analysis for the purpose of restoration and do not covers the speed requirements for large scale archive quality analysis. A young field of research is how calculated measures are corresponding with human perception, this key research area has to be followed.

18.4 Content Analysis Tools for Audio and Speech

Many methods and descriptors using different sets of low-level audio features have been proposed in literature during the last years. Many of them represent the efforts toward the answer to the emerging needs of tools for automatic segmentation, classification, identification and retrieval of audio material. For each situation where audio content analysis tools could provide a valuable improvement, an appropriate choice of low-level features and analysis methods must be always considered.

18.4.1 Segmentation/classification

Segmentation and classification are mandatory for split audio streams into homogeneous regions according to the desiderate criteria. These methods can offer a precious time saving support both for human and for automatic activities that otherwise have to deal with the whole content of an audio stream. The computational power required by segmentation/classification methods is quite light, easily allowing real-time processing.

18.4.2 Pattern retrieval

Pattern retrieval could represent an important method for accessing to audio archives using similarity matching (query by example paradigm) or simply for tracking well known events (like jingles, signature tunes or other patterns). The latter case requires a mandatory training phase where a set of known events must be supplied.

18.4.3 Automatic speech recognition

A different application domain is represented by automatic speech recognition, where a relevant synergy between phonetic, linguistic and statistic domains achieve impressive results (e.g. more than 90% of correct words with planned speech in controlled acoustic environment). In order to obtain the above high scores from an automatic speech recognition process it is mandatory to have a linguistic and a lexical model representing the considered application domain, otherwise an important loss of performance (measured in word error rate) can be experienced. It should be obvious that both lexical and linguistic models are representative then usable only for each specific language. Although today computer's computational power is impressive and less expensive than few years ago, an automatic speech recognition process still require a huge computational power (a real-time speech recognition process for an audio stream requires at least a cluster of 3 bi-processor@2.8 GHz).

18.5 Joint Audiovisual Content Analysis Tools

18.5.1 Scene/Story Segmentation

Scene and story segmentation requires a definition of the unit to be segmented. As a scene is not the same the same of all types of material and there is no single valid definition, many approaches focus on news material, as a news story can be defined unambiguously. There are however attempts to find more generic scene definitions and approaches which are not based on heuristically derived domain knowledge.

The results of the TRECVID 2004 news story segmentation task show, that precision and recall up to 0.7-0.8 can be reached for news material. There is a clear indication that the results are better if visual and audio information is combined, however, additional use of automatic speech recognition results does not significantly improve the results.

An evaluation method for parsing video content into segments has been proposed, which measures the gain for the user in terms on manual work, comparing the effort needed to manually segment the content with effort needed to correct an automatic segmentation. The gain values are between 77% and 95% [Hanj04].

18.5.2 Shot and Scene Classification

There are a number of approaches for genre classification, ranging from scene to programme level. Most of them use some prior knowledge of the genres involved, so that there is no useful generic approach. There are however algorithms that work well with a limited set of sufficiently discriminative classes (e.g. commercial, non-commercial).

The TRECVID 2003 and 2004 results show that the precision for some classification tasks is good enough for practical use (e.g. weather news, sporting event). The reason is that they can be linked to certain low-level features, e.g. weather news has characteristic colours, sports events have high motion energy and a high background noise level. For classes where this relation is not so well defined, the precision is very poor (e.g. detection of beach scenes).

18.5.3 Event Detection

Many event detection approaches are based on domain specific prior knowledge. A large number of these approaches focus on event detection in sports video, as they have a clear structure and the events of interest are known a priori. However, some researchers have proposed generalised

approaches to event detection, although a part of those also needs training or domain knowledge to form rules for event inference.

Specialised event detection approaches yield better results than generic ones. The performance is generally moderate and depends very much on how well visual and audio features correlate with the event.

18.5.4 Video Content Abstraction

There are two classes of approaches for video content abstraction: video summarisation (representing video content using collections of still images, which are sometimes navigatable) and video skimming (produces short, trailer-like videos form longer content).

In the form of sets of key frames, video summarisation is commonly used. But event his simple approach suffers from the fact that all summaries and skims are context dependent (e.g. getting overview over material, judging relevance of a search result) and sometimes user dependent. The main contradiction in selecting the material to be included in a summary is between proportionally representing the original content and selecting the outstanding events, that may just cover a small fraction of the content.

Video content abstraction depends crucially on the input it can use. It can benefit from high-level content analysis results, especially from scene and story segmentation, classification and event detection, as only the results of these analysis steps allow automatically determining the relevance of a part of the material for a summary in a certain context.

19 Dependencies between CA Tools

This section consists mainly of a graph that visualises the dependencies between the audiovisual content analysis tools, from low-level to high level. The high-level tools typically describe features, which are of high practical relevance for the user. The intention of this graph is to visualise, which low- and mid-level content analysis tools are required as prerequisites for the extraction of meaningful high-level metadata.

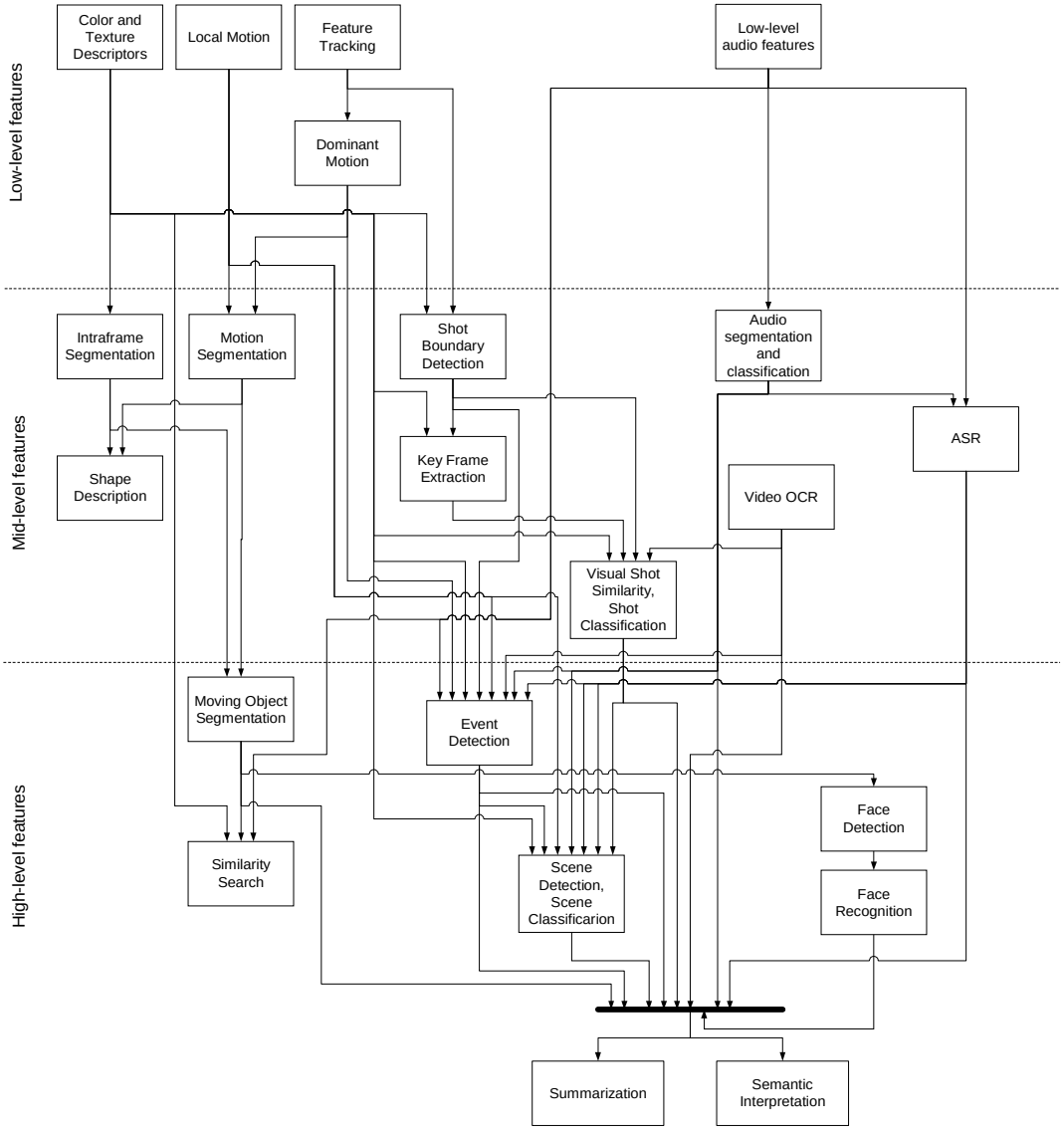


Figure 8: Dependencies between content analysis tools.

Appendix

20 References

[AAW01] A. A. Alatan, A. N. Akansu and W. Wolf, “Multi-modal dialog scene detection using Hidden Markov Models for content based multimedia indexing”, *Multimedia Tools and Applications*, 14(14):137–151, 2001.

[Ada03] N. Adami, R. Leonardi, P. Migliorati, “An Overview of Multi-modal Techniques for the Characterization of Sport Programmes”, *Proc. Conf. Visual Communications and Image Processing*, Lugano, Jul, 2003, pp. 1296-1306.

[AH00] S. Aksoy, R. M. Haralick, “Using texture in image similarity and retrieval”, Texture Analysis in Machine Vision, Pietikainen Ed. Vol. 20, pp. 129-149, 2000.

[Ala02] A. A. Alatan, “Automatic multi-modal dialogue scene indexing”, Technical report, Electrical-Electronics Engineering Dept., Univ. Ankara, Turkey, 2002.

[Amir04] A. Amir et al., “IBM Research TRECVID-2004 Video Retrieval System”, Proc. TRECVID workshop 2004.

[Anan93] P. Anandan, J. R. Bergen, K. J. Hanna and R. Hingorani, “Hierarchical Model-Based Motion Estimation”, In: M. I. Sezan and R. L. Lagendijk (eds.), *Motion analysis and image sequence processing*, Kluwer Academic Publishers, 1993, pp. 1—22.

[ANSI801-96] American National Standards Institute, *Digital Transport of One-Way Video Signals - Parameters for Objective Performance Assessment*, ANSI T1.801.03-1996.

[Arm94] F. Arman, R. Depommier, A. Hsu and M. Y. Chiu, “Content-based browsing of video sequences”, ACM Multimedia’94, pp. 97-103, Aug. 1994.

[Bai02] W. Bailer, *Motion Estimation and Segmentation for Film and Video Standards Conversion and Restoration*, Diploma Thesis, Fachhochschul-Diplomstudiengang Media Technology and Design, Hagenberg, Austria, 2002.

[BDbP99] S. Berretti, A. Del Bimbo, P.Pala, "Retrieval by shape Using Multidimensional Indexing Structures", Proceedings of the 10th ICIAP, Venezia, 1999, pp. 945-950.

[Ber04] M. Bertini, R. Cucchiara, A. Del Bimbo and A. Prati, ”Objects and Events Recognition for Sport Videos Transcoding”, *Proc. International Symposium on Image/Video Communications*, Brest, Jul. 2004.

[Bess04] B. Besserer and C. Thire. “Detection and Tracking Scheme for Line Scratch Removal in an Image Sequence”. In The 8th European Conference on Computer Vision - ECCV 2004.

[Beu91] S. Beucher, “The Watershed Transformation applied to image segmentation”. 10th Pfefferkorn Conf. on Signal and Image Processing in Microscopy and Microanalysis, 16-19 sept. 1991, Cambridge, UK. In: Scanning Microscopy International, suppl. 6. 1992, pp. 299-314.

[BIM97] A. Del Bimbo, P. Pala, “Visual Image Retrieval by Elastic Matching of User Sketches”. IEEE, 1997.

[Boch00] L. Boch, "TV Programmes Acquisition System for RAI Multimedia Catalogue", Elettronica e Telecomunicazioni, Anno XLIX, nr. 1, pp. 38-47, 2000. (*in italian*)

[Bor96] J. S. Boreczky and L. A. Rowe, “Comparison of Video Shot Boundary Detection Techniques,” In: Storage and Retrieval for Still Image and Video Databases IV, Proc. SPIE 2664, pp. 170-179, Jan. 1996.

[Bovik00] A. C. Bovik and S. Liu, “DCT-domain blind measurement of blocking artifacts in DCT-coded images,” Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc., vol. 3, pp. 1725-1728, May 2001.

[Bra96] C.J. van den Branden Lambrecht, “A working spatio-temporal model of the human visual system for image restoration and quality assessment applications”. In Conference Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 1996.

[Brava] The Brava broadcast archive programme impairments dictionary. URL: http://brava.ina.fr/brava_public_impairments_list.en.html.

[Buis03] O. Buisson, S. Boukir, and B. Besserer. “Motion compensated film restoration”. Machine Vision and Applications, 13(4):pp. 201-212, Feb. 2003.

[Camp99] P. Campisi, A. Longari and A. Neri, “Automatic key frame selection using a wavelet based approach”, Proc. of SPIE, vol. 3813, pp. 861-872, July 1999.

[Cam98] G. Campra, "Digital processing of television signals for multimedia cataloguing", Tesi di Laurea, Politecnico di Torino, 1998.

[Cav00] J. E. Caviedes, A. Drouot, A. Gesnot, and L. Rouvellou, "Impairment metrics for digital video and their role in objective quality assessment," Proc. SPIE, vol. 4067, pp. 791-800, 2000.

[Cav02] J. Caviedes, S. Gurbuz, “No-reference sharpness metric based on local edge kurtosis”. In proceedings of ICIP 2002, vol. 3, pp.53-56.

[Cettolo04] Mauro Cettolo, Michele Vescovi, Romeo Rizzi “Evaluation of BIC-based algorithms for audio segmentation” July 2004.

[Cha05] M. Chambah, C. Saint Jean, F. Helt, “Image Quality Evaluation in the field of digital film restoration”. In: Image Quality and System Performance II, Proc. SPIE 5668, 2005.

[Cha97-1] M. M. Chang, A. M. Tekalp, and M. I. Sezan. Simultaneous motion estimation and segmentation. IEEE Trans. Image Proc., 6(9):1326–1333, Sept. 1997.

[Cha97-2] S. Chang, W. Chen, H. J. Meng , H. Sundaram , D. Zhong, VideoQ, *Proceedings of the fifth ACM international conference on Multimedia,* November 1997

[Chai02] L. Chaisorn and T.-S. Chua, “The Segmentation and Classification of Story Boundaries In News Video”, *Proc. of 6th IFIP working conference on Visual Database Systems- VDB6 2002*, Australia 2002.

[Chan04] H.-C. Chang and S.-H. Lai, “Robust camera motion estimation and classification for video analysis”, *Proc. SPIE-IS&T Visual Communications and Image Processing*, San Jose, CA, Jan. 2004.

[Chen04] M. Chen and J. Yang, “Feature Extraction Techniques. CMU at TRECVID 2004”, Proc. TRECVID workshop 2004.

[Clark00] P. Clark and M. Mirmehdi, “Finding Text Regions Using Localised Measures.”, *Proceedings of the 11th British Machine Vision Conference,* 2000, pp. 675-684.

[Clausi02] D. A. Clausi, “An analysis of co-occurence texture statistics as a function of grey level quantisation”, Canadian Journal of Remote Sensing, Vol. 28 nr. 1, pp 45-62, 2002.

[CLPO99] L. Cinque, S. Levialdi, A. Pellicanò, K. A. Olsen, “Colour-based image retrieval using spatial chromatic histograms”, IEEE Transactions on Multimedia Systems 1999, Vol. 2 pp.969-973.

[Com97] D. Comaniciu, P. Meer, “Robust analysis of feature spaces: Colour image segmentation”. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, San Juan, Puerto Rico, June 1997, pp. 750-755.

[Cou01] Francois-Xavier Coudoux, Marc Georges Gazalet, Christian Derviaux, and Patrick Corlay, “Picture quality measurement based on block visibility in discrete cosine transform coded video sequences”. Journal of Electronic Imaging, April 2001, Volume 10, Issue 2, pp. 498-510.

[CRG97] C. Colombo, A. Rizzi, I. Genovesi, “Histogram families for colour based retrieval in image databases”, ICIAP ’97 Proceedings, Florence Sept. 1st 1997, pp.204-211.

[CRÖ03] L. Chen, J. Rizvi, and M. T. Özsu, “Incorporating audio cues into dialog and action scene extraction”, Technical report, School of Computer Science, University of Waterloo, Waterloo, Canada, 2003.

[CriST00] N. Cristianini, J Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press 2000 ISBN: 0 521 78019 5

[Dang95] Viet-Nam Dang, Abdol-Reza Mansouri, and Janusz Konrad. Motion estimation for region-based video coding. In Proceedings of IEEE International Conference on Image Processing, pages 189–192, Washington, DC, Oct. 1995.

[DAV97] E. R. Davies, *Machine Vision: Theory, Algorithms, Practicalities*. Academic Press, 1997.

[Davy02] M. Davy, S.J. Godsill “Audio Infromation retrieval: a Bibliographic Study” , Cambridge University, February 2002

[Dbim99] A. Del Bimbo, *Visual Information Retrieval*, Morgan Kaufmann Publishers Inc. 1999.

[Dem98] D. Dementhon, V. Kobla and D. Doermann, “Video summarization by curve simplification”, ACM Multimedia 98, pp. 211-218, 1998.

[DEN02] Dengsheng Zhang, *Image Retrieval Based on Shape*. Dissertation, Faculty of Information Technology at Monash University, 2002.

[Den99] Y. Deng, B.S. Manjunath and H. Shin, “Colour image segmentation”. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Fort Collins, CO, vol.2, Jun. 1999, pp.446-51.

[Dim00] N. Dimitrova, L. Agnihotri, and G. Wei, “Video classification based on HMM using text and faces,” in European Signal Processing Conference, Tampere, Finland, 2000.

[Dim03] N. Dimitrova, “Multimedia content analysis: The next wave”, *Proc. International Conference on Image and Video Retrieval*, Urbana-Champaign, IL, USA, Jul. 2003, pp. 9-18.

[Div03] Divakaran et al, "Video Summarization Using MPEG-7 Motion Activity and Audio Descriptors", Video Mining, Rosenfeld, A.; Doermann, D.; DeMenthon, D., October 2003 (Kluwer Academic Publishers), http://www.merl.com/reports/docs/TR2003-34.pdf

[Dor04] A. Dorado, D. Djordjevic and E. Izquierdo, “Supervised semantic scene classification based on low-level clustering and relevance feedback”, *Proc. European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies*, London, Nov. 2004.

[Doul00] A. D. Doulamis, N. Doulamis and S. Kollias, “Non-sequential video content representation using temporal variation of feature vectors”, IEEE transactions on Consumer Electronics, vol. 46, no. 3, August 2000.

[Dub93] E. Dubois and J. Konrad, “Estimation of 2-D Motion Fields from Image Sequences with Application to Motion-Compensated Processing”, In: M. I. Sezan and R. L. Lagendijk (eds.), *Motion analysis and image sequence processing*, Kluwer Academic Publishers, 1993, pp. 53—87.

[Duda01] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, 2nd ed, Wiley & Sons, 2001.

[Duf95] F. Dufaux , F. Moscheni, and A. Lippman. Spatio-temporal segmentation based on motion and static segmentation. In Proceedings of IEEE International Conference on Image Processing, pages 306–309, Washington, DC, Oct. 1995.

[Duf96] F. Dufaux and Fabrice Moscheni. Segmentation-based motion estimation for second generation video coding techniques. In L. Torres and M. Kunt, editors, Video Coding: The Second Generation Approach, chapter 6, pages 219–263. Kluwer Academic Publishers, 1996.

[Duf00] F. Dufaux, “Key frame selection to represent a video”, ICME2000.

[Eich04] D. Eichmann and D.-J. Park, “Boundary and Feature Extraction at the University of Iowa”, Proc. TRECVID workshop 2004.

[Eng96] P. England, R. B. Allen, M. Sullivan and A. heybey, “I/Browse: The bellcore video library toolkit”, Proc. of SPIE, vol. 2670, pp. 254-264, Feb. 1996.

[EvalFaceR] Evaluation of Face Recognition Alogrithms, URL: http://www.cs.colostate.edu/evalfacerec.

[EzC02] S. Ezekiel, J. A. Cross, "Fractal-based texture analysis”, Proceedings of ACM SIGCSE 2002, Cincinnati 2002.

[Far05] M.Q.Farras, J.M. Foley, S.K. Mitra, Univ. of California, “Perceptual analysis of video impairments that combine blocky, blurry, noisy, and ringing synthetic artefacts”. In: Human Vision and Electronic Imaging X, Proc. SPIE 5666, 2005.

[Fis95] S. Fischer, R. Lienhart, and W. Effelsberg, “Automatic recognition of film genres,” *Proc. ACM Multimedia 1995*, San Francisco, USA, 1995, pp. 295–304.

[Fis03] Dublin City University (centre for digital image processing), Fischlar-TV, <URL: http://www.cdvp.dcu.ie>

[Flor00] L. Florak, "A spatio-frequency trade-off scale for scale space filtering", IEEE Trans. On Pattern Analysis and Machine Intelligence, Vol. 22 n° 9, pp. 1050-1055, 2000.

[Foley90] J. Foley, A. van Dam, S. Feiner, and J. Hughes. Computer Graphics: Principles and Practice. Systems Programming Series. Addison-Wesley, second edition, 1990.

[FRE61] H. Freeman, “On the Encoding of Arbitrary Geometric Configurations”, IRE Trans. on Electronic Computers, Vol. EC-10, 1961.

[Frisch] R. Frischholz, “The Face Detection Home Page”, URL: http://home.t-online.de/home/Robert.Frischholz/face.htm.

[FRTV1a] Report on FRTV Phase I. Available: ftp://ftp.crc.ca/crc/vqeg/phase1-docs/ final_report_april00.pdf.

[FRTV1b] Report on FRTV Phase I. Available: ftp://ftp.its.bldrdoc.gov/dist/ituvidq/phase1_final_report/COM-80E.pdf.

[FRTV2] Report on FRTV Phase II. Available: ftp://ftp.its.bldrdoc.gov/dist/ituvidq/frtv2_final_report/ VQEGII_Final_Report.pdf.

[FUS74] K. S. Fu, *Syntactic Methods in Pattern Recognition*, Academic Press, 1974.

[Gast01] P. Gastaldo, S. Rovetta and R. Zunino, “Objective assessment of MPEG-video quality: a neural-network approach,” in Proc. IJCNN, vol. 2, pp. 1432-1437, 2001.

[Gerb02] Ralf Gerber et al, "Deriving Textual Descriptions of Road Traffic Queues from Video Sequences", Proceedings 15th European Conference on Artificial Intelligence, Lyon, France, July 2002, S. 736-740, http://cogvisys.iaks.uni-karlsruhe.de/publications/rg_hhn_hs_ECAI2002.pdf

[Gil04] W. J. Gillespie and D. T. Nguyen, "Classification of Video Shots using Activity Power Flow", *Proc. of 2004 IEEE Consumer Communications and Networking Conference*, Las Vegas, USA, 2004.

[Gish91] H.Gish, M.Siu, R.Rohlicek, “Segregation of speakers for speech recognition and speaker identification” in Proc. ICASSP – Toronto, Canada (1991).

[GMMMO03] S. Guha, A. Meyerson, N. Mishra, R. Motwani, L. O'Callaghan, "Clustering Data Streams: Theory and Practice", IEEE Transactions on Data and Knowledge Engineering, Vol. 15, nr. 3, pp 515-528, 2003.

[Goh04] K.-S. Goh, K. Miyahara, R. Radhakrishan, Z. Xiong, A. Divakaran, “Audio-Visual Event Detection based on Mining of Semantic Audio-Visual Labels”, *Proc. Conf. Storage and Retrieval for Multimedia*, 2004.

[GOS85] A. Goshtasby, “Description and Discrimination of Planar Shapes Using Shape Matrices”, IEEE, 1985.

[GRO92] W. I. Groskey, P. Neo, R. Mehrotra, “A Pictorial Index Mechanism for Model-Based Matching”, Data & Knowledge Engineering, 1992.

[Guy02] N. Guyader, H. Le Borgne, J. Hérault an A. Guérin-Dugué, “Towards the introduction of human perception in a natural scene classification system”, *Proc. IEEE Workshop on Neural Networks for Signal Processing*, Sept. 2002.

[Hab04] G. Haberfehlner, *Development of a System for Automatic Dialog Scene Detection*, Diploma Thesis, Fachhochschul-Diplomstudiengang Software Engineering, Hagenberg, Austria, 2004.

[Hae00] N. Haering, R. J. Qian and M. I. Sezan, “A Semantic Event-Detection Approach and Its Application To Detecting Hunts in Wildlife Video*”, IEEE Trans. Circuits and Systems for Vidoe Technology*, vol. 10, nr. 6, Sep. 2000.

[Hae01] N. Hearing and N. da Vitoria Lobo, *Visual Event Detection*, Kluwer Academic Publishers, 2001.

[Hanj97] A. Hanjalic, M. Ceccarelli, R. L. Lagendijk, and J. Biemond, “Automation enabling search on stored video data”, Proc. of SPIE, vol. 3022, pp. 427-438, 1997.

[Hanj99a] A. Hanjalic and H. J. Zhang, “An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis”, IEEE Transactions on Circuits and Systems for Video Technology, vol. 9, no. 8, Dec. 1999.

[Hanj99b] A. Hanjalic, R. L. Lagendijk and J. Biemond, “Automated High-Level Movie Segmentation for Advanced Video-Retrieval Systems”, *IEEE Trans. On Circuits and Systems for Video Technology*, vol. 9, nr. 4, Jun. 1999, pp. 580-588.

[Hanj00] A. Hanjalic, G. C. Langelaar, P. M. B. Van Roosmalen, J. Biemond and R. L. Lagendijk, *Image and Video Databases: Restoration, Watermarking and Retrieval*, Elsevier, 2000.

[Hanj02] A. Hanjalic, “Shot-boundary detection: unraveled and resolved?” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 2, pp. 90 -105, Feb. 2002.

[Hanj04] A. Hanjalic, *Content-based analysis of digital video*, Kluwer Academic Publishers, 2004.

[Haup98] A. Hauptmann and M. J. Witbrock, “Story Segmentation and Detection of Commercials In Broadcast News Video”, *Advances in Digital Libraries*, 1998.

[Hei86] G. W. Heiman, R. J. Leo, G. Leighbody and K. Bowler, “Word intelligibility decrements and the comprehension of time-compressed speech”, Perception and psychophysics, vol. 40, no. 6, pp. 407-411, 1986.

[Heesch04] D. Heesch et al., ”Video Retrieval using Search and Browsing”, Proc. TRECVID workshop 2004.

[Hjelmas01] Erik Hjelmas and Boon Kee Low, “Face Detection: A Survey”, *Computer Vision and Image Understanding,* 2001, pp. 236-274.

[Hoas04] K. Hoashi et al., “Shot Boundary Determination on MPEG Compressed Domain and Story Segmentation Experiments for TRECVID 2004”, Proc. TRECVID workshop 2004.

[Horn81] B. Horn and B. Schunck, “Determining optical flow”, *Artifical Intelligence*, vol 17, 1981, pp. 185—203.

[Hsu04] W. Hsu et al., “Discovery and Fusion of Salient Multi-modal Features towards News Story Segmentation”, Proc. Storage and Retrieval Methods and Applications for Multimedia, San Jose, CA, 2004, pp. 244–258.

[HUM62] M. K. Hu. Visual Pattern Recognition by Moment Invariants. IRE Transactions on Information Theory, 1962.

[Hunt87] R. W. G. Hunt, *The Reproduction of Colour in Photography, Printing and Television*. 4th ed., Fountain Press, 1987.

[Hunt98] F.E.T. Huntsberger, B.D. Jawerth, T. Kubota, "Wavelet-Based Fractal Signature Analysis for Automatic Target Recognition", Optical Engineering 37(1) 166-174, Jan. 1998.

[Iane03] T. I. Ianeva, A. P. de Vries, A.P. and H. Rohrig, “Detecting cartoons: a case study in automatic video-genre classification*”. Proc. International Conference on Multimedia and Expo*, 2003.

[ITU-T_J.143] International Telecommunications Union, Telecommunication Standardization Sector, "User requirements for objective perceptual video quality measurements in digital cable television", ITU-T Recommendation J.143.

[JAP00] Jaroslav Pokorny, “Prostorove datove struktury a jejich pouziti k indexaci prostorovych objektu”, Katedra softwaroveho inzenyrstvi, MFF UK, 2000.

[Jin04] S. H. Jin, T. M. Bae,, J. H. Choo and Y. M. Ro, “Video Genre Classificatin Using Multimodal Features”, *Proc. Storage and Retrieval for Multimedia*, San Jose, CA, USA, Jan. 2004.

[JMT03] G. Joubert, S. M. Mouattamid, H. H. Tadjine, "Colour Object Detection with Shadow Using Virtual Electric Field", ICISP 2003.

[Johson99] S.E. Johnson, P. Woodland “Speaker clustering using direct maximisation of the MLLR-adapted likelihood” 1999, Cambridge University.

[Joy00] Laurent Joyeux, Samia Boukir, Bernard Besserer. „Film Line Scratch Removal Using Kalman Filtering and Bayesian Restoration”. WACV’2000, IEEE Workshop on the Application of Computer Vision, Dec. 2000.

[Ju98] S. X. Ju, M. J. Black, S. Minneman and D. Kimber, “Summarization of video-taped presentations: automatic analysis of motion and gestures”, IEEE Transactions on CSVT, 1998.

[Kang03] Y.-L. Kang, J.-H. Lim, Q. Tian and M. S. Kankanhalli, “Soccer Video Event Detection with Visual Keywords”, *Proc. ICICS-PCM*, Singapore, 2003, pp. 1796-1800.

[Kemp00] T.Kemp, M. Schmidt, M. Westpal, A. Waibel, “Strategies for automatic segmentation of audio data“ In Proc. ICASSP – Istanbul, Turkey (2000).

[Kim02] Jung-Rim Kim et al "Scalable Hierarchical Video Summarization of News using Fidelity in MPEG-7 Description Scheme," Lecture Notes in Computer Science, VISUAL 2002, pp.239-246 , March 2002 http://mpeg.korea.ac.kr/paper/Summarization_Published.pdf

[Knee01] M. Knee, “A robust, efficient and accurate single-ended picture quality measure for MPEG-2,” available at http://www-ext.crc.ca/vqeg/frames.html, 2001.

[Koho97] T. Kohonen, "Self-Organizating Maps", Springer-Verlag NY 1997.

[Kop01] I. Koprinska and S. Carrato, “Temporal Video Segmentation: A Survey,” Signal Processing: Image Communication, vol. 16, pp. 477–500, 2001.

[Kok98] Anil Kokaram. "Motion Picture Restoration - Digital Algorithms for Artefact Suppression in Degraded Motion Picture Film and Video”. Springer Verlag 1998, ISBN 3-540-76040-7.

[Kraa04] W. Kraaij and J. Arlandis, “TRECVID-2004 Story segmentation task: Overview”, Proc. TRECVID workshop 2004.

[Krü96] S. Krüger and A. Calway. Multiresolution motion estimation using an affine model. Technical Report CSTR-96-002, University of Bristol, 1996.

[KWT98] M. Kass, A. Witkin, D. Terzopoulos, "Snakes: active contour models", International Journal of Computer Vision, Vol. 1 nr. 4 pp. 321-331.

[Lag96] R. L. Lagendijk, A. Hanjalic, M. Ceccarelli, M. Soletic and E. Persoon, “Visual search in a smash system”, Proc. of ICIP’96, pp. 671-674, Sep. 1996.

[Laws80] K, Laws, “Rapid texture identification”, SPIE vol. 238, Image Processing for Missile Guidance, pp. 376-380.

[Leh04] B. Lehane, N. E. O'Connor and N. Murphy, "Action Sequence Detection in Motion Pictures", *Proc. EWIMT*, London, Nov. 2004.

[Lee02] B. R. Lee, A. Ben Hamza and Hamid Krim, "An Active Contour Model for Image Segmentation: A Variational Perspective", IEEE Internat. Conf. on Acoustics Speech and Signal Processing, Orlando, Florida, May 2002.

[Leo04] R. Leonardi, P. Migliorati and M. Prandini, "Semantic Indexing of Soccer Audio-Visual Sequences: A Multimodal Apporach Based on Controlled Markov Chains", *IEEE Trans. Circuits and Systems for Video Technology*, vol. 14, nr. 5, May 2004.

[Leung95] T.K. Leung, M.C. Burl, and P. Perona, "Finding faces in cluttered scenes using random labelled graph matching", *Proc. Of The Fith International Conference on Computer Vision,* 1995.

[Li03] Y. Li and C.-C. Jay Kuo, *Video Content Analysis Using Multimodal Information*, Kluwer Academic Publishers, 2003.

[Li95] S. Z. Li. Markov Random Field Modeling in Computer Vision. Springer, New York, 1995.

[Lien00] R. Lienhart, "Dynamic video summarization of home video", Proc. of IS&T/SPIE, vol. 3972, pp. 378-389, Jan. 2000.

[Lien01] R. Lienhart, "Reliable Transition Detection In Videos: A Survey and Practitioner's Guide," International Journal of Image and Graphics (IJIG), vol. 1, No. 3, pp. 469-486, 2001.

[Lienhart03] Rainer Lienhart, "Video OCR: A Survey and Practitioner´s Guide.", *The Kluwer International Series in Video Computing, Volume 6: Video Mining,* 2003.

[LP94] F. Liu, R. Picard, "A new Wold ordering for image similarity" Proceedings of the IEEE conference on Acoustics, Speech and Signal Processing, Adelaide (AUS) 1994, pp. 129-132.

[LP96] F. Liu, R. Picard, "Periodicity, directionality and randomness: Wold features for image modelling and retrieval", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18 pp 722-733, 1996.

[Lu98] C.-S. Lu and P.-C. Chung, "Wold Features for Unsupervised Texture Segmentation", Proc. 14th IAPR Inter. Conf. on Pattern Recognition, vol. II, pp. 1689-1693, 1998.

[Lu02] Lie Lu, Hong Zhang, "Content Analysis for Audio Classification and segmentation", *IEEE Transaction on speech and audio processing*, vol 10. no 7. October 2002.

[Lu04a] Shi Lu et al "Video summarization by spatial-temporal graph optimization", In Proceedings of IEEE ISCAS 2004, 2004, pages II--197--200, Vancouver, Canada, May 2004. IEEE Society. http://www.cse.cuhk.edu.hk/~king/PUB/iscas2004_cd.pdf

[Lu04b] Shi Lu et al "Video summarization by video structure analysis and graph optimization", In Proceedings of the 2004 IEEE International Conference on Multimedia and Expo (ICME'04), volume CD-ROM, Taipei, Taiwan, June2004. IEEE Society. http://www.cse.cuhk.edu.hk/~king/PUB/icme2004_cd.pdf

[Luc81] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision", Proc. DARPA image understanding workshop, 1981, pp. 121—130.

[m4] m4 - multimodal meeting manager, http://www.m4project.org/

[Manj01] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, A. Yamada, "Colour and Texture Descriptors", *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 11, no. 6, Jun. 2001.

[Mich93] R.S. Michalski, J.W. Bala, P.W. Pachowicz, "GMU Research in learning in vision: initial results", Proceedings of the DARPA Image Understanding Workshop, April 1993, Washington.

[Mills92] M. Mills, "A magnifier tool for video data", Proc. of ACM Human Computer Interface, pp. 93-98, May 1992.

[Molau01] Sirko Molau, Michael Pitz, et al “Computing Mel-Frequency Cepstral Coefficient on the Power Spectrum” 2001, Aachen, University of Technology, Germany.

[Moon96] T. K. Moon. The expectation-maximization algorithm. IEEE Signal Processing Magazine, 13(6):47–60, November 1996.

[MM97] W.Y. Ma, B. S. Manjunath, “NeTra: a toolbox for navigating large image databases”, ICIP ’97 Proceedings, Santa Barbara CA, pp. 568-571.

[Mont04] M. Montagnuolo, “Tools for automatic classification of audiovisual content”, Tesi di Laurea, Politecnico di Torino, Sept. 2004 (in italian).

[MPEG7-3] ISO/IEC 15938-3: *Information Technology - Multimedia Content Description Interface*, Part 3 Visual, 2001.

[MPEG7-5] ISO/IEC 15938-5: *Information Technology - Multimedia Content Description Interface*, Part 5 Multimedia Description Schemes, 2001.

[Nam98] J. Nam, M. Alghoniemy, and A.H. Tewfik, “Audio-visual content-based violent scene characterization,” in IEEE International Conference on Image Processing, Chicago, USA, 1998, Vol. 1, pp. 353–357.

[Naph03] M. Naphade, J. R. Smith, “A Hybrid Framework for Detecting the Semantics of Concepts and Context“, *Proc. 2nd Intl. Conf. Image and Video Retrieval*, pp. 196-205, Urbana-Champaign, IL, 2003.

[Nash97] J. M. Nash, J. N. Carter, and M. S. Nixon. Dynamic feature extraction via the velocity hough transform. Pattern Recognition Letters, 18(10):1035–1047, 1997.

[Nitt02] N. Nitta, N. Babaguchi and T. Kitahashi, “Story Based Representation for Broadcasted Sports Video and Automatic Story Segmentation”, *Proc. IEEE International Conference on Multimedia and Expo (ICME2002)*, pp.813-816, 2002.

[OCon01] N. O’Connor, C. Czirjek, S. Deasy, S. Marlow, N. Murphy and A. Smeaton, “News Story Segmentation in the Físchlár Video Indexing System”, *Proc. ICIP 2001 - International Conference on Image Processing*, Thessaloniki, Greece, 10-12 October 2001.

[Ojala02] T. Ojala, M. Aittola, E. Matinmikko, “Empirical evaluation of MPEG-7 XM colour descriptors in content-based retrieval of semantic image categories”, *Proc. 16th International Conference on Pattern Recognition*, vol. 2, pp. 1021 – 1024, Aug. 2002.

[Omo99] N. Omoigui, L. He, A. Gupta, J. Grudin and E. Sanocki, “Time-compression: System concerns, usage, and benefits”, Proc. of ACM Conference on Computer Human Interaction, 1999.

[OTT91] P. J. van Otterloo, *A Contour-Oriented Approach to Shape Analysis*. Prentice Hall International (UK), 1991.

[Pan01] H. Pan, P. van Beek, M. I. Sezan, “Detection of slow-motion replay segments in sports video for highlights generation”, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001, pp. 1649-1652.

[Pap04] D. M. Papworth, E. Izquierdo and A. Pearman, “Using HMMs to classify Video Sequences”, *Proc. European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies*, London, Nov. 2004.

[Pet04] C. Petersohn, “Fraunhofer HHI at TRECVID 2004: Shot boundary detection system”, Proc. TRECVID workshop 2004.

[Piat00] J.H. Piater, R. A. Grupen, "Constructive Feature Learning and the Development of Visual Expertise", Proceedings of the 17th International Conference on Machine Learning, Stanford, CA, June 29-July 2, 2000. Morgan Kaufmann.

[Pick03] M. J. Pickering, L. Wong and S. M. Rüger, “ANSES: Summarisation of news video”, *Proc. of International Conference on Image and Video Retrieval (CIVR)*, Urbana-Chmpaign, IL, USA, July 2003.

[PLE99] S. Pfeiffer, R. Lienhart and W. Effelsberg, “Scene determination based on video and audio features”, Technical report, University of Mannheim, Praktische Informatik IV, 68131 Mannheim, Germany, 1999.

[Pre92] W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, *Numerical Recipes in C*, Second edition, Cambridge University Press, 1992.

[PZM96] G. Pass, R. Zabih, J. Miller , “Comparing images using colour coherence vectors”, Proceedings of the ACM Conference on Multimedia, Boston MA 1996, pp. 65-73.

[Que04] G. M. Quénot et al., “CLIPS-LIS-LSR-LABRI experiments at TRECVID 2004”, Proc. TRECVID workshop 2004.

[Rash02] Z. Rasheed and M. Shah, “Movie genre classification by exploiting audio-visual features of previews”, *Proc. International Conference on Pattern Recognition*, Aug. 2002.

[Rash03] Z. Rasheed and M. Shah, “Scene Detection In Hollywood Movies and TV Shows”, *Proc. Conf. on Computer Vision and Pattern Recognition*, Madison, WI, USA, June 2003.

[Rata99] K. Ratakonda, M. I. Sezan and R. Crinon, “Hierarchical video summarization”, Proc. Of SPIE, vol. 3653, pp. 1531-1541, Jan. 1999.

[Rau03] M. Rautianen, T. Seppanen, J. Pentilla and J. Peltola, “Detecting semantic concepts from video using temporal gradients and audio classification”, Proc. Conf. on Image and Video Retrieval, Urbana-Champaign, USA, Jul. 2003.

[RBK98] H. A. Rowley, S. Baluja, and T. Kanade, “Neural network-based face detection”, Technical report, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA, 1998.

[Rui00] Y. Rui and P. Anandan, “Segmenting visual actions based on spatio-temporal motion patterns”, *Proc. Conf. Computer Vision and Pattern Recognition*, 2000.

[Rui99] Y. Rui, T. S. Huang and S. Mehrotra, “Constructing Table-of-Content for Videos”, *ACM Multimedia Systems Journal*, Sept 1999.

[Russ03] Daniel M. Russell, Andreas Dieberger "Synthesizing evocative imagery through design patterns", In HICSS'36 (Hawaii International Conference on Systems Science), Waikaloa, HI, January 2003, http://homepage.mac.com/juggle5/WORK/publications/HICSS2003.pdf

[Russ00] D. M. Russell, “A design pattern-based video summarization technique: moving from low-level signals to high-level structure”, Proc. of the 33rd Hawaii International Conference on System Sciences, vol. 1, Jan. 2000.

[Sad04] D. Sadlier, N. O'Connor, N. Murphy and S. Marlow, “A Framework for Event Detection in Field-Sports Video Broadcasts based on SVM generated Audio-Visual Feature Model. Case-Study:Soccer Video”, *International Workshop on Systems, Signals and Image Processing,* Poznan, Poland, Sept. 2004.

[Scha02] F. Schaffalitzky and A. Zisserman, “Automated scene matching in movies”, Proc. Conf. on Image and Video Retrieval, London, UK, Jul. 2002, pp. 176-185.

[Schall99] P. Schallauer, A. Pinz, W. Haas. *Automatic Restoration Algorithms for 35mm Film.* Videre: Journal of Computer Vision Research, Vol. 1, Number 3, pp. 60-85, Summer 1999, MIT Press. WWW: http://www-mitpress.mit.edu/e-journals/Videre/001/v13.html

[Schall02] P. Schallauer, M. Donoser, G. Kienast, H. Rehatschek, “High Quality Compression for Film and Video”, PRESTO Deliverable D5.4, 2002.

[Schwarz78] G.Schwarz “Estimating the dimension of a model”, The Annual of statistics, 6,2 (1978).

[SchemaD21] M. Barlaud et al., State of the art in content-based analysis, indexing and retrieval, IST-2001-32795 D2.1, Sep. 2002. Available: http://www.iti.gr/SCHEMA/preview.html?file_id=67.

[SD97] M. Stricker, A. Dimai, "Spectral covariance and fuzzy regions for image indexing“, Machine Vision and Applications, Vol. 10, pp. 66-73, 1997.

[SEK92] I. Sekita, T. Turita, N. Otsu, “Complex Autoregresive Model for Shape Recognition”, IEEE, 1992.

[Ser02] N. Serrano, A. E. Savakis, and J. Luo, “A computationally efficient approach to indoor/outdoor scene classification”, Proc. Intl. Conf. Pattern Recognition, 2002, vol. IV, pp. 146-149.

[Shi00] J. Shi and J. Malik, “Normalized Cuts and Image Segmentation”. IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 22, no. 8, Aug. 2000.

[SiMa04] S. Singh, M. Markou, "An approach to novelty detection Applied to the Classification of Image Regions", IEEE Transactions on Data and Knowledge Engineering, Vol. 16 Nr. 4 pp. 396-407, April 2004.

[SL99] C. Saraceno and R. Leonardi, “Identification of story units in audio visual sequences by joint audio and video processing”, *Proc. Int´l Workshop on Multimedia Signal Processing*, pp. 53–58, 1999.

[Smea02] A. F. Smeaton: "Challenges for content-based navigation of digital video in the Fischlar digital library", proceedings of CIVR 2002 conference, pp.: 215–224.

[Snoek04] C. G. M. Snoek, M. Worring, J. M. Geusebroek, D. C. Kolma and F. J. Seinstra, “The MediaMill TRECVID 2004 Semantic Video Search Engine”, Proc. TRECVID workshop 2004.

[Snoek05] C.G.M. Snoek and M. Worring, “Multimodal Video Indexing: A Review of the State-of-the-art”, *Multimedia Tools and Applications*, 25(1):5-35, Jan. 2005.

[SON93] M. Sonka, V. Hlavac, R. Boyle, *Image Processing, Analysis and Machine Vision*. Chapman & Hall Computing, 1993.

[SOTV04] A. Smeaton and P. Over, “Shot Boundary Detection Task Overview”, Proc. TRECVID workshop 2004.

[Souv04] F. Souvannavong, B. Merialdo and B. Huet, “Eurécom at Video-TREC 2004: Feature Extraction Task”, Proc. TRECVID workshop 2004.

[SPSV01] M. De Santo, G. Percannella, C. Sansone, and M. Vento, “Dialogue scenes detection in mpeg movies: A multi-expert approach”, Technical report, Dipartimento di Ingegneria dell Informazione e di Ingegneria Elettrica Universita di Salerno, Via P.te Don Melillo,1 I-84084, Fisciano (SA), Italy, 2001.

[Still99] C. Stiller and J. Konrad, “Estimating motion in image sequences. A tutorial on modelling and computation of 2D motion”, *IEEE Signal Processing Magazine*, vol 16, nr. 4, July 1999, pp. 70—91.

[Smi97] M. A. Smith and T. Kanade, “Video skimming and characterization through the combination of image and language understanding techniques”, Proc. of the IEEE Computer Vision and Pattern Recognition, pp. 775-781, 1997.

[Stef00] A. Stefanidis, P. Partsinevelos, P. Agouris and P. Doucette, “Summarizing video datasets in the spatiotemporal domain”, Proc. of 11th International Workshop on Database and Expert Systems Applications, pp. 906-912, Sep. 2000.

[Sto04] M. Štochl, *Object Indexing and Matching for Video Retrieval*, Diploma Thesis, Dept. of Cybernetics, Univ. of West Bohemia, Pilsen, 2004.

[Strehl00] Alexander Strehl and J. K. Aggarwal. A new Bayesian relaxation framework for the estimation and segmentation of multiple motions. In Proceedings of the 4th IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI 2000), pages 21–25, April 2000.

[Sug03] M. Sugano, R. Isaksson, Y. Nakajima and H. Yanagihara, “Shot genre classification using compressed audio-visual features”, *Proc. International Conference on Image Processin*g, Sep. 2003.

[Sun03] H. Sun, J-H. Lim, Q. Tian and M. S. Kankanhalli, “Semantic Labelling of Soccer Video”, *Proc. ICICS-PCM*, Singapore, Dec. 2003.

[Sun00] X. D. Sun and M. S. Kankanhalli, “Video summarization using R-sequences”, Real-time Imaging, pp. 449-459, Dec. 2000.

[Sund02] H. Sundaram and S.-F. Chang, “Computable Scenes and Structures in Films”, *IEEE Transactions on Multimedia*, vol. 4, nr. 4 , Dec. 2002, pp. 482-491.

[Szum98] M. Szummer and R.W. Picard, “Indoor-outdoor image classification,” in IEEE International Workshop on Content-based Access of Image and Video Databases, in conjunction with ICCV’98, Bombay, India,1998.

[Tan00] Y.-P. Tan, D. D. Saur, S. R. Kulkarni and P. J. Ramadge, “Rapid Estimation of Camera Motion from Compressed Video with Application to Video Annotation”, *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 10, nr 1, Feb. 2000, pp. 133—146.

[TAU92] G. Taubin, D. B. Cooper, “Object Recognition Based on Moment (or Algebraic) Invariants”, In: J. Mundy and A. Zisserman (eds.): *Geometric Invariance in Computer Vision*, 1992.

[Tava04] W. Tavanapong and J. Zhou, “Shot Clustering Techniques for Story Browsing”, *IEEE Trans. on Multimedia*, vol. 6, nr. 4, Aug. 2004, pp. 517-527.

[TEA80] Michael Reed Teague, “Image Analysis Via the General Theory of Moments”, Journal of Optical Society of America, 1980.

[Tek95] A. Murat Tekalp. Digital video processing. Prentice Hall, 1995.

[TextSeg] Text Localisation/Text Segmentation. URL: http://www.videoanalysis.org/Research_Topics/Text_Localization__Text_Segmen/text_localization__text_segmen.html.

[Thang00] Thang V. Pham, Marcel Worring, “Face Detection Methods: A Critical Evaluation”, *ISIS Report*, 2000.

[Tok00] C. Toklu, A. P. liou and M. Das, “Videoabstract: A hybrid approach to generate semantically meaningful video summaries”, ICME2000, New York, 2000.

[Tov01] V. Tovinkere and R. J. Qian, “Detecting Semantic Events in Soccer Games: Towards a Complete Solution”, *Proc. IEEE Conf. On Multimedia and Expo*, 2001.

[TP91] M. Turk and A. Pentland, “Eigenfaces for recognition”, *Journal of Cognitive Neuroscience*, pages 71–86, March 1991.

[Trecvid] TREC Video Retrieval Evaluation. URL: http://www-nlpir.nist.gov/projects/trecvid

[Truo00] B. T. Truong and C. Dorai, “Automatic genre identification for content-based video categorization”, *Proc. International Conference on Pattern Recognition*, Sep. 2000.

[Truo02] B. T. Truong, S. Venkatesh and C. Dorai. “Film Grammar Based Refinements to Extracting Scenes in Motion Pictures”, *IEEE International Conference on Multimedia and Expo (ICME2002)*, pp- 281-284.

[Tsai01] Yaakov Tsaig and Amir Averbuch. Automatic segmentation of moving objects in video sequences: A region labeling approach. IEEE Transactions on Circuits and Systems for Video Technology, 2001.

[TVKA04] W. Kraaj and J. Arlandis, “Feature extraction task: Overview”, Proc. TRECVID workshop 2004.

[Uchi99] S. Uchihashi, J. Foote, A. Girgensohn and J. Boreczky, “Video manga: generating semantically meaningful video summaries”, ACM Multimedia’99, 1999.

[Ueda93] H. Ueda, T. Miyatake, S. Sumino and A. Nagasaka, “Automatic structure visualization for video editing”, Proc. of INTERCHI’93, pp. 137-141, 1993.

[VA03] http://www.videoanalysis.org/Research_Topics/Video_Segmentation/Shot_Detection/shot_detection.html

[Vail98] A. Vailaya, A.K. Jain, and H.-J. Zhang, “On image classification: City images vs. landscapes,” Pattern Recognition, Vol. 31, pp. 1921–1936, 1998.

[Vail00] A. Vailaya and A.K. Jain, “Detecting sky and vegetation in outdoor images,” in Proceedings of SPIE: Storage and Retrieval for Image and Video Databases VIII, San Jose, USA, 2000, Vol. 3972.

[Vasc98] N. Vasconcelos, A. Lippman, “A spatiotemporal motion model for video summarization”, Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 1998.

[Velt02] R. C. Veltkamp, M. Tanase, “Content-Based Image Retrieval Systems: A Survey”, Technical Report UU-CS-2000-34, Department of Computer Science, Utrecht University. Revised and extended version, October 2002.

[Vend02] J. Vendrig and M. Worring, “Systematic Evaluation of Logical Story Unit Segmentation”, *IEEE Trans. on Multimedia*, vol. 4, nr. 4, Dec. 2002, pp. 492-498.

[Vin91] L. Vincent and P. Soile, “Watersheds in digital spaces: an efficient algorithm based on immersion simulations”. IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 13, nr. 6, pp. 583-597, June 1991.

[Vin93] L. Vincent, “Morphological greyscale reconstruction in image analysis: efficient algorithms and applications”. IEEE Trans. on Image Processing, pp. 176-201, 1993.

[Vla04] T. Vlachos, “Flicker correction for archived film sequences using a non-linear model”, IEEE Trans. on Circ. and Sys. for Video Tech., Vol. 14, No. 4, 508-516, April 2004.

[Volk04] T. Volkmer, S. M. M. Tahahoghi and H. E. Williams, “RMIT University at TRECVID 2004”, Proc. TRECVID workshop 2004.

[VQEG] Video Quality Experts Group. URL: http://www.vqeg.org.

[Wa93] John Y. A. Wang and Edward H. Adelson. Layered representation for motion analysis. In Proceedings of the IEEE Computer Vision and Pattern Recognition Conference, pages 361–366, New York, Jun. 1993.

[Wa94] John Y. A. Wang and Edward H. Adelson. Representing moving images with layers. IEEE Trans. Image Proc., 3(5):625–638, Sept. 1994.

[Wang00a] Y. Wang, Z. Liu, and J. Huang, “Multimedia content analysis using both audio and visual clues,” *IEEE Signal Processing Magazine*, Vol. 17, No. 6, pp. 12–36, 2000.

[Wang00b] Wang, A. C. Bovik and B. L. Evans, “Blind measurement of blocking artifacts in images,” Proc. IEEE Int. Conf. Image Proc., vol. 3, pp. 981-984, Sept. 2000.

[Wang04] J. Wang, C. Xu, E. S. Chng and Q. Tian, “Sports Highlight Detection from Keyword Sequences Using HMM”, *Proc. International Conference on Multimedia and Expo*, 2004.

[Wei97a] Yair Weiss. Motion segmentation using EM - a short tutorial. URL, http://www.cs.berkeley.edu/~yweiss/emTutorial.ps, 1997.

[Wei97b] Yair Weiss, “Smoothness in Layers: Motion segmentation using non-parametric mixture estimation”, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 1997, pp. 520—527.

[Whit04] A. Whitehead, P. Bose and R. Laganiere, “Feature Based Cut Detection with Automatic Threshold Selection”, Proc. Intl. Conf. Image and Video Retrieval, Dublin, 2004, pp. 411–418.

[Win95] G. Winkler, *Image analysis, random fields and dynamic Monte Carlo methods: A mathematical introduction.*, Springer, 1995.

[Wolf96] W. Wolf, “Key frame selection by motion analysis”, ICASSP’96, vol. 2, pp. 1228-1231, 1996.

[Wu97] H. R. Wu and M. Yuen, “A generalized block-edge impairment metric for video coding,” IEEE Signal Processing Letters, vol. 4, pp. 317-320, Nov. 1997.

[Wu01] P. Wu, Y. M. Ro, C. S. Won, Y. Choi, "Texture Descriptors in MPEG-7", *Proc. 9th International Conference Computer Analysis of Images and Patterns (CAIP)*, LNCS 2124, p.21-28, Warsaw, Sept. 2001.

[Xin02] Xin Li, “Blind image quality assessment”, in Proceedings on ICIP 2002, vol. 1, pp. 449-452.

[Xio03] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. S. Huang, "Audio Events Detection Based Highlights Extraction from Baseball, Golf and Soccer Games in a Unified Framework", *Proc. IEEE International Conference on Multimedia and Expo*, Jul. 2003, pp. 401-404.

[XP97] C. Xu, J.L. Prince, "Gradient Vector Flow: A New External Force for Snakes" Proc. IEEE Conf. on Comp. Vis. Patt. Recog. (CVPR), Los Alamitos: Comp. Soc. Press, pp. 66-71, June 1997.

[YAN98] H. S. Yang, S. U. Lee, K. M. Lee, “Recognition of 2D Object Contours Using Starting-Point-Independent Wavelet Coefficient Matching”, Journal of Visual Communication and Image Representation, 1998.

[Yang02] M. H. Yang, D. Kriegman, and N. Ahuja. “Detecting faces in images: A Survey”, *IEEE TRANSACTION PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 2002.

[Yang04] M. H. Yang, “Face Detection Methods: A Critical Evaluation”, *ISIS Report*, 2000.

[Ying01] Ying Li et.al., "An Overview of Video Abstraction Techniques", HP Laboratories Palo Alto HPL-2001-191, 2001

[Ying03] Ying Li, Kuo, C.C. Jay, "Video Content Analysis Using Multimodal Information For Movie Content Extraction, Indexing and Representation" 2003, 224 p., ISBN: 1-4020-7490-5

[Yow96] K.C. Yow, R. Cipolla, “Scale and orientation invariance in human face detection”, *British Machine Vision Conference,* 1996, pp. 745-754.

[Yu97] H. Yu, G. Bozdagi and S. Harrington, ”Feature based hierarchical video segmentation,” Proc. ICIP'97, Santa Barbara, Sept. 1997.

[Yuan04] J. Yuan et al., “Tsinghua University at TRECVID 2004: Shot boundary detection and high-level feature extraction”, Proc. TRECVID workshop 2004.

[YYL96] M. Yeung, B. L. Yeo, B. Liu, "Extracting story units from long programmes for video browsing and navigation", Proceedings of the International Conference on Multimedia Computing and Systems, Hiroshima 1996.

[Zel01] L. Zelnik-Manor and M. Irani, "Event-Based Analysis of Video*", Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Dec. 2001.

[ZEL03] M. Zelezny, “Zpracovani digitalizovaneho obrazu”, Lectures, Faculty of Applied Sciences, Univ. West Bohemia, 2003 (in Czech).

[Zhai04] Y. Zhai et al., “University of Central Florida at TRECVID 2004”, Proc. TRECVID workshop 2004.

[Zhang97] H. J. Zhang, J. Wu, D. Zhong and S. W. Smoliar, “An integrated system for content based video retrieval and browsing”, pattern Recognition, vol. 30, no. 4, pp. 643-658, 1997.

[Zhang98a] Tong Zhang and C.-C. Jay Kuo, “Content Based Classfication and Retrieval of Audio”, 1998.

[Zhang98b] T.Zhang, C. Kuo, “Hierarchical system for content-based audio classification and retrieval“, 1998.

[Zhao03] W. Zhao, R. Chellappa, and P.J. Phillips, “Face Recognition: A Literature Survey”, *ACM Computing Surveys, Vol. 35, No. 4,* 2003, pp. 399-458.

[Zhong01] D. Zhong and S.-F. Chang, “Structure Analysis of Sports Video Using Domain Models”, *IEEE Conference on Multimedia and Exhibition*, Tokyo, Japan, Aug. 22-25, 2001.

21 Glossary

Term	Definition
ASR	Automatic speech recognition, a.k.a. speech to text.
Camera motion	The motion caused by camera movement and zoom.
CIE Luv	A colour space defined by CIE (Commission Internationale de l'Éclairage/International Commission on Illumination), that is perceptually uniform, i.e. the Euclidian distance between two colours in the colour space corresponds to the distance perceived by humans.
CSS	Curvature scale space, a method for the description of the contour of a region.
cut	A cut is a point in an aural or visual event at which the perceived content changes abruptly. Audio and video feature extraction algorithms generally detect cuts by tracking abrupt changes in the measurement of a set of selected features.
DCT	Acronym for Discrete Cosine Transform, a mathematical transformation representing a discrete two-dimensional signal data as a weighted sum of cosines.
Descriptor	(1) A set for data that compactly represents a feature. (2) In MPEG-7, a descriptor is an atomic unit of a metadata description. Sets of related descriptors form description schemes.
DFT	Short for Discrete Fourier Transform, a mathematical transformation applied to a discrete complex valued series obtaining a series of complex values representing the frequency decomposition of a digital signal.
Dominant motion	The motion of the background or of an object in the scene.
DPD	Displaced Pixel Difference: the greyscale or colour difference between a frame at time t and the prediction of the frame using estimated motion and projecting frame t-1 or t+1 to t.
Feature	A property derived from a part of audiovisual data that can be used to describe and represent the AV data.
FFT	Acronym for Fast Fourier Transform, an optimised algorithm for calculating the DFT of a signal which number of samples is a power of 2.
Global motion	The motion to which the whole image is subject to (“background motion”). Moving objects move relatively to the global motion.
GMM	Gaussian Mixture Model. Statistical modelling technique where different pdf (probability distribution functions) are jointly describe using a mean vector and a covariance matrix. This technique is a powerful tool for speaker modelling.
GoF/GoP	Group of Frames/Group of Pictures. A group of frames is a generic ordered set of contiguous frames in a video sequence. A group of pictures is a group of frames selected as the basic repetitive structure of the MPEG2 video coding algorithm.
HMM	Hidden Markov Model. Very simplistically, a hidden Markov model is based on a set of probabilistic functions (state transition distributions and symbol observation probabilities) used to calculate the probabilities of determined observation symbols sequences out of a modelled system.

Term	Definition
HSV	A colour space that describes colour in terms of the three components Hue, Saturation and Value (Lightness).
Local motion	Motion calculated for blocks, regions or small pixel patches, that only takes the direct neighbourhood into account. The result is typically a motion vector field, that describes the displacement of every pixel or block of pixels.
Low-level feature	A visual or audio feature that can be directly extracted from the signal, not using any further prior knowledge. Low-level features are colour, texture, shape and motion.
MFCC	Mel Frequency Cepstral Coefficients, a low-level audio feature (cf. Section 11.8)
MPEG-7	MPEG-7 (ISO/IEC 15938, formally named “Multimedia Content Description Interface”) is a standard for describing multimedia content, independent of the encoding of the content, and allows different levels of granularity of the description. MPEG-7 has been designed to support a broad range of applications. MPEG-7 descriptions can be represented either as XML (textual format, TeM) or in a binary format (binary format, BiM).
MRF	Markov Random Field, a field with Markovian properties (i.e. the state of an element in the field is only determined by its neighbours).
OCR	Optical Character Recognition
PCA	Principal Components Analysis is a technique that can be used to simplify a dataset [Duda01].
PDE	Partial Differential Equation
Shot	A shot is a sequence of fundamental units of a audio or video document resulting from an uninterrupted sensor recording.
Shot boundary	A cut or an editing effect (e.g. fade, dissolve, wipe) that delineates two shots.
SVD	Singular Value Decomposition, a method for solving linear algebraic equations, commonly used to solve linear least-squares problems [Pre92].
SVM	Acronym for Support Vector Machine. A support vector machine is a data clustering method based on a non linear kernel-based transformation of a feature space into another space where transformed data vectors can be separated by means of hyperplanes defined in the transformed space. SVMs are trained by means of support vectors, i.e. a set of feature space data which probabilistic distribution is the same as that of the total feature space.
Texture	The structure of an image region.
Transition	A shot boundary. Cuts are called abrupt transitions, editing effects such as fades, dissolves and wipes are called gradual transitions.
VQEG	Video Quality Experts Group
VQM	Video Quality Measurement